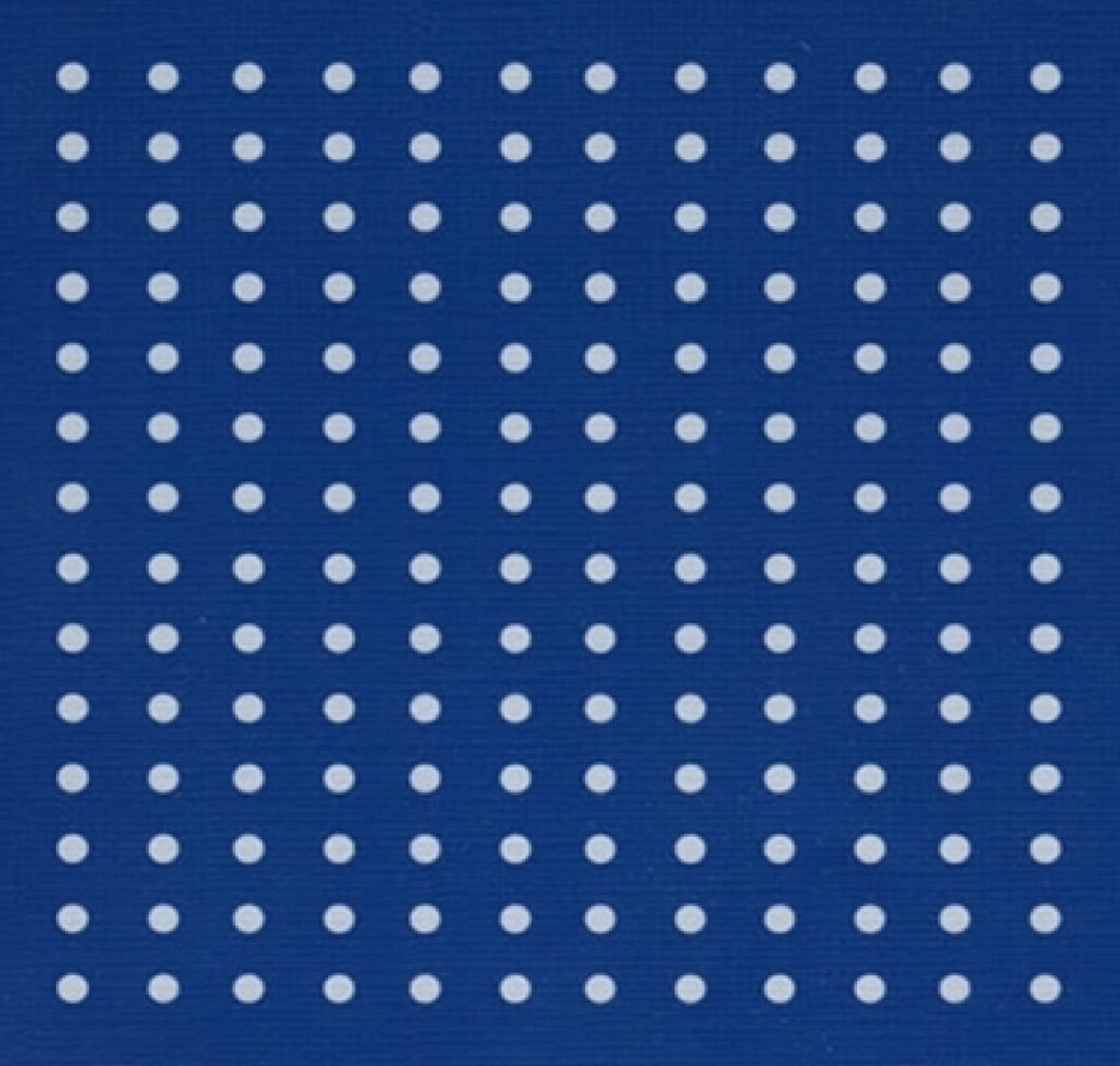
重点大学计算机专业系列教材

数据挖掘原理与算法（第二版）教师用书

毛国君 段立娟 编著



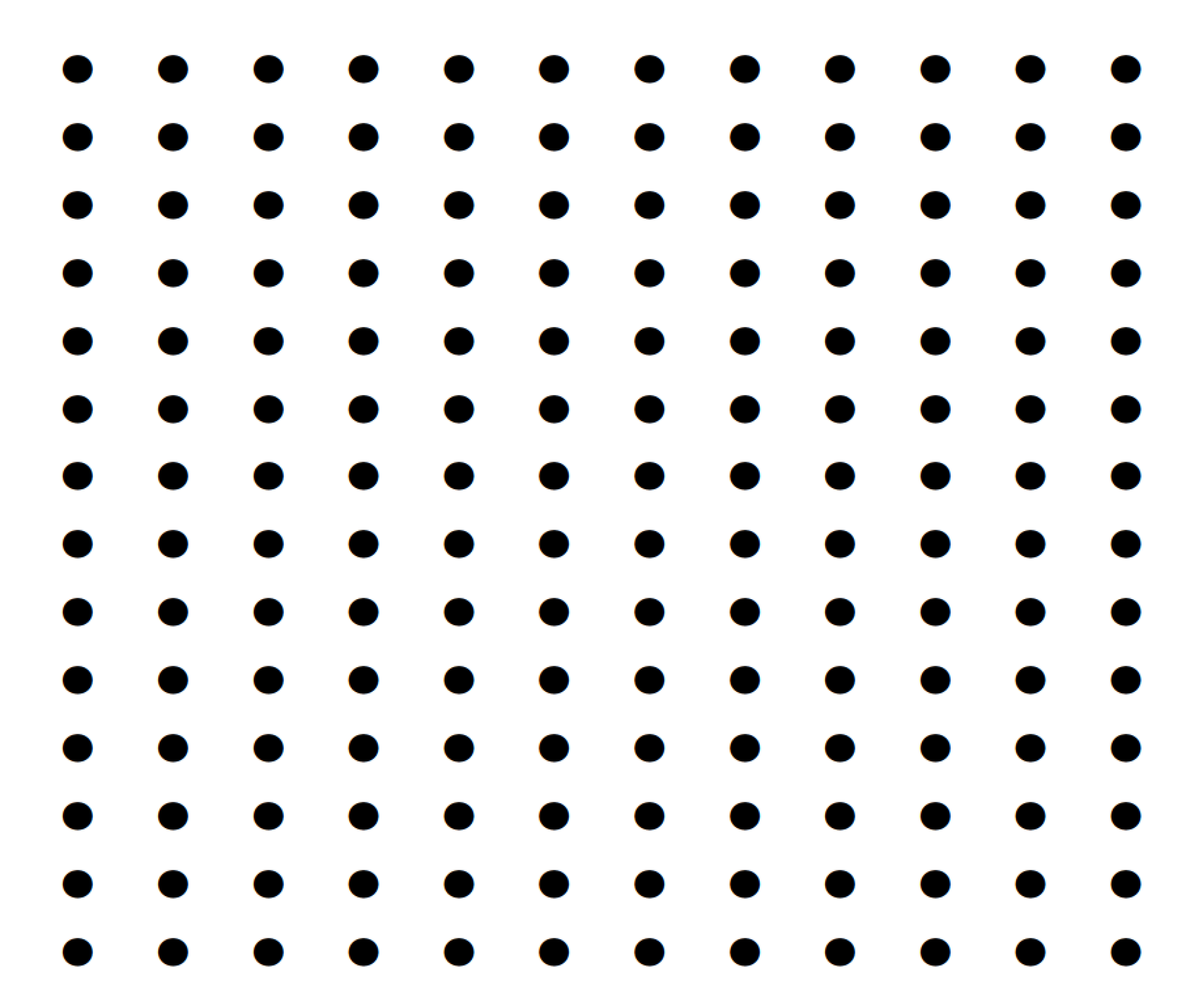
清华大学出版社



普通高校“十一五”规划教材
重点大学计算机专业系列教材

数据挖掘原理与算法（第二版）教师用书

毛国君 段立娟 编著



清 华 大 学 出 版 社
北 京

内 容 简 介

《数据挖掘原理与算法》一书出版以来，被许多高校作为本科生或者研究生教材使用，是一本全面介绍数据挖掘和知识发现技术的专业书籍，具有内容系统、知识含量高等特点。为了让教师更好地使用教材《数据挖掘原理与算法》(第二版)，作者又编写了本书。本书分四个部分：一、对教材每章的部分习题给出了参考答案；二、介绍各章授课内容重点与课时分配；三、针对不同的授课学生对象给出了课时安排的建议；四、提供了两套样本试卷及其参考答案。

本书供使用《数据挖掘原理与算法》一书的教师作参考。

图书在版编目(CIP)数据

数据挖掘原理与算法(第2版)教师用书/毛国君，段立娟编著. —2版. —北京：清华大学出版社，2009.3
(重点大学计算机专业系列教材)
ISBN 978-7-302-19350-0

Ⅰ.数… Ⅱ.①毛… ②段… Ⅲ.数据采集—高等学校—教学参考资料 Ⅳ.TP274

中国版本图书馆CIP数据核字(2009)第010846号

责任编辑：丁 岭 张为民
责任校对：焦丽丽
责任印制：

出版发行：清华大学出版社 **地 址**：北京清华大学学研大厦A座
http://www.tup.com.cn **邮 编**：100084
社 总 机：010-62770175 **邮 购**：010-62786544
投稿与读者服务：010-62776969，c-service@tup.tsinghua.edu.cn
质 量 反 馈：010-62772015，zhiliang@tup.tsinghua.edu.cn
印 刷 者：
装 订 者：
经 销：全国新华书店
开 本：185×260 **印 张**：5.5 **字 数**：134千字
版 次：2009年2月第1版 **印 次**：2009年2月第1次印刷
印 数：1～ 000
定 价： .00元

本书如存在文字不清、漏印、缺页、倒页、脱页等印装质量问题，请与清华大学出版社出版部联系调换。联系电话：010-62770177转3103 产品编号：

INTRODUCTION

出版说明

随着国家信息化步伐的加快和高等教育规模的扩大，社会对计算机专业人才的需求不仅体现在数量的增加上，而且体现在质量要求的提高上，培养具有研究和实践能力的高层次的计算机专业人才已成为许多重点大学计算机专业教育的主要目标。目前，我国共有16个国家重点学科、20个博士点一级学科、28个博士点二级学科集中在教育部部属重点大学，这些高校在计算机教学和科研方面具有一定优势，并且大多以国际著名大学计算机教育为参照系，具有系统完善的教学课程体系、教学实验体系、教学质量保证体系和人才培养评估体系等综合体系，形成了培养一流人才的教学和科研环境。

重点大学计算机学科的教学与科研氛围是培养一流计算机人才的基础，其中专业教材的使用和建设则是这种氛围的重要组成部分，一批具有学科方向特色优势的计算机专业教材作为各重点大学的重点建设项目成果得到肯定。为了展示和发扬各重点大学在计算机专业教育上的优势，特别是专业教材建设上的优势，同时配合各重点大学的计算机学科建设和专业课程教学需要，在教育部相关教学指导委员会专家的建议和各重点大学的大力支持下，清华大学出版社规划并出版本系列教材。本系列教材的建设旨在“汇聚学科精英、引领学科建设、培育专业英才”，同时以教材示范各重点大学的优秀教学理念、教学方法、教学手段和教学内容等。

本系列教材在规划过程中体现了如下一些基本组织原则和特点。

1. 面向学科发展的前沿，适应当前社会对计算机专业高级人才的培养需求。教材内容以基本理论为基础，反映基本理论和原理的综合应用，重视实践和应用环节。

2. 反映教学需要，促进教学发展。教材要能适应多样化的教学需要，正确把握教学内容和课程体系的改革方向。在选择教材内容和编写体系时注意体现素质教育、创新能力与实践能力的培养，为学生知识、能力、素质协调发展创造条件。

3. 实施精品战略，突出重点，保证质量。规划教材建设的重点依然是专业基础课和专业主干课；特别注意选择并安排了一部分原来基础比较好的

优秀教材或讲义修订再版，逐步形成精品教材；提倡并鼓励编写体现重点大学计算机专业教学内容和课程体系改革成果的教材。

4. 主张一纲多本，合理配套。专业基础课和专业主干课教材要配套，同一门课程可以有多本具有不同内容特点的教材。处理好教材统一性与多样化的关系；基本教材与辅助教材以及教学参考书的关系；文字教材与软件教材的关系，实现教材系列资源配套。

5. 依靠专家，择优落实。在制订教材规划时要依靠各课程专家在调查研究本课程教材建设现状的基础上提出规划选题。在落实主编人选时，要引入竞争机制，通过申报、评审确定主编。书稿完成后要认真实行审稿程序，确保出书质量。

繁荣教材出版事业，提高教材质量的关键是教师。建立一支高水平的以老带新的教材编写队伍才能保证教材的编写质量，希望有志于教材建设的教师能够加入到我们的编写队伍中来。

教材编委会

FOREWORD

前言

《数据挖掘原理与算法》一书出版以来，被许多高校作为本科生或者研究生的教材使用。几年来许多教师给出了很好的建议，因此我们在 2007 年针对相关问题进行了修订并出版了其第二版。该教材是一本全面介绍数据挖掘和知识发现技术的专业书籍，具有内容系统、知识含量高等特点。可能也正是因为这些特点，作为教材给教师带来了一些授课难点。特别是，由于教材使用的对象不同，对教材内容进行选择是必需的。为了让教师更好地使用《数据挖掘原理与算法》一书，减轻教师的负担，我们编写了本教师用书。

《数据挖掘原理与算法(第二版)教师用书》主要从四个部分为教师提供了参考：一、对教材每章的部分习题给出了参考答案；二、介绍各章授课内容重点与课时分配；三、针对不同的授课学生对象给出了课时安排的建议；四、提供了两套样本试卷及其参考答案。

目的是为了帮助教师提高讲课的效率，但不能代替教师的教学研究工作。特别考虑到教师用书也可能被学生使用，故对教材后面的习题并没有给出全部解答。

整体上说，数据挖掘技术包含概念与过程、原理与方法两个主要部分。对于有关概念与过程，主要集中在《数据挖掘原理与算法》(第二版)第 1 章和第 2 章，不论学生对象如何，教师都应该给予重视，力求全面而直观地进行介绍。数据挖掘中的原理与方法，分布在《数据挖掘原理与算法》(第二版)的第 3~ 8 章，涵盖关联规则、分类、聚类、序列、空间以及 Web 挖掘等分支。我们认为，关联规则、分类、聚类是经典内容，不论学生对象如何，教师都应该选择一些典型的理论和算法进行剖析。对于不同的教学对象，教师可以对第 3~ 5 章的内容进行合理选择。例如，如果准备给本科生开一个只有 32 课时的课程，那么最起码的要求是在对于关联规则、分类、聚类等基本概念和原理讲述清楚的前提下，能把 Apriori、ID3 和 k-means 算法剖析清楚即可。第 6~ 8 章的内容相对比较松散，对于研究生来说，我们认为需要进行选择性地介绍或讨论。这是因为这些内容属于数据挖掘的较前沿的课题，而且有着很广泛的研究和应用价值，因此对于研究生将来的研究工作可能会有很大的帮助。

《数据挖掘原理与算法》(第二版)共分 8 章，各章相对独立，而且每章的

内容都是从前往后难度逐渐增大的。因此,教师完全可以发挥自己的想象力和知识上的优势进行内容选择。此外,如果读者是从事计算机相关研究和开发的人员,本教师用书可能也能帮助读者节约宝贵时间,提高《数据挖掘原理与算法》(第二版)一书的利用效率。总之,作者希望通过本教师用书,提供一个很好地利用《数据挖掘原理与算法》(第二版)的辅助材料,促进数据挖掘技术的普及与提高。

作　者

2008 年 12 月于北京

CONTENTS

目录

P A R T 1

第一部分 各章习题及部分参考答案

第1章　绪　　论

1. 给出下列英文缩写或短语的中文名称和简单的含义

（1）Data Mining

（2）Artificial Intelligence

（3）Machine Learning

（4）Knowledge Engineering

（5）Information Retrieval

（6）Data Visualization

参考答案：（1）数据挖掘。简单地说就是从大型数据中挖掘所需要的知识。

（2）人工智能。简单地说就是研究如何应用机器来模拟人类某些智能行为的基本理论、方法和技术的一门科学。

（3）机器学习。简单地说就是研究如何使用机器来模拟人类学习活动的一门学科。

（4）知识工程。简单地说就是研究知识信息处理并探讨开发知识系统的技术。

（5）信息检索。简单地说就是研究合适的信息组织并根据用户需求快速而准确地查找信息的技术。通常指的是计算机信息检索，它以计算机技术为手段，完成电子信息的汇集、存储和查找等的相关技术。

（6）数据可视化。简单地说就是运用计算机图形学和图像处理等技术，将数据换为图形或图像在屏幕上显示出来。它是进行人机交互处理、数据解释以及提高系统可用性的重要手段。

2. 给出下列英文缩写或短语的中文名称和简单的含义。

（1）OLTP（On-line Transaction Processing）

（2）OLAP（On-line Analytic Processing）

（3）Decision Support

（4）KDD（Knowledge Discovery in Databases）

（5）Transaction Database

（6）Distributed Database

参考答案：略。

3. 为什么说数据挖掘是未来信息处理的骨干技术之一？

参考答案：数据挖掘之所以被称为未来信息处理的骨干技术之一，主要在于它以一种全新的概念改变着人类利用数据的方式。数据挖掘和知识发现使数据处理技术进入了一个更高级的阶段。它不仅能对过去的数据进行简单地查询，并且能够找出过去数据之间的潜在联系，进行更高层次的分析，以便更好地做出理想的决策、预测未来的发展趋势等。

4. 从商业需求角度分析数据挖掘技术产生的合理性。

参考答案：略。

5. 支撑数据挖掘技术的主要研究基础学科有哪些？说明数据挖掘产生的技术背景。

参考答案：任何技术的产生总是有它的技术背景的。数据挖掘技术的提出和普遍接受

是由于计算机及其相关技术的发展为其提供了研究和应用的技术基础。普遍认为,对数据挖掘产生决定性作用的三个主要技术是:数据库技术、统计学和包括机器学习在内的人工智能技术。

在关系型数据库的研究和产品提升过程中,人们一直在探索组织大型数据和快速访问的相关技术。高性能关系数据库引擎以及相关的分布式查询、并发控制等技术的使用,已经提升了数据库的应用能力。在数据的快速访问、集成与抽取等问题的解决上积累了经验。数据仓库作为一种新型的数据存储和处理手段,被数据库厂商普遍接受并且相关辅助建模和管理工具快速推向市场,成为多数据源集成的一种有效的技术支撑环境。因此,人们已经具备利用多种方式存储海量数据的能力。这些丰富多彩的数据存储、管理以及访问技术的发展,为数据挖掘技术的研究和应用提供了丰富的土壤。

计算机芯片技术的发展,使计算机的处理和存储能力日益提高。随之而来的是硬盘、CPU 等关键部件的价格大幅度下降,使得人们收集、存储和处理数据的能力和欲望不断提高。经过几十年的发展,计算机的体系结构,特别是并行处理技术已经逐渐成熟和普遍应用,并成为支持大型数据处理应用的基础。计算机性能的提高和先进的体系结构的发展使数据挖掘技术的研究和应用成为可能。

历经了十几年的发展,包括基于统计学、人工智能等在内的理论与技术性成果已经被成功地应用到商业处理和分析中。这些应用从某种程度上为数据挖掘技术的提出和发展起到了极大地推动作用。数据挖掘系统的核心模块技术和算法都离不开这些理论和技术的支持。从某种意义上讲,这些理论本身的发展和应用为数据挖掘提供了有价值的理论和应用积累。

6. 数据挖掘技术是一个交叉研究分支,简述影响它产生和发展的主要研究学科或分支及其关系。

参考答案:略。

7. 数据(Data)、信息(Information)和知识(Knowledge)是人们认识和利用数据的三个不同阶段,数据挖掘技术是如何把它们有机的结合在一起的?

参考答案:从数据、信息和知识三个层面上看,数据是最原始的未经组织和处理的信息源。信息或称有效信息是指对人们在某些方面有价值的东西。知识是一种现实世界信息的抽象和浓缩,是一种概念、规则、模式和规律等。数据挖掘技术通过对原始数据进行微观、中观乃至宏观的统计、分析、综合和推理,发现数据间的关联性、未来趋势以及一般性的概括知识等,转变成可以用来指导人们某些高级商务活动的有用信息。

8. 从数据挖掘研究角度看,如何理解数据、信息和知识的不同和联系。

参考答案:略。

9. 简述数据挖掘技术将来的发展趋势。

参考答案:对于数据挖掘技术的发展趋势,应该分两方面辩证的理解。

(1) 数据挖掘技术已经存在相当大市场,将成为对工业产生重要影响的关键技术之一。同时,并行计算机体系结构研究和 KDD 也被列入今后 5 年内公司应该投资的 10 个新技术领域之一。这些资料都表明,数据挖掘技术在将来有很大的发展潜力及空间。

(2) 数据挖掘技术作为一门新技术,仍有许多问题需要研究、解决和探索。分析目前的研究和应用现状,对于数据挖掘技术将来的工作重点有:

① 数据挖掘技术与特定商业逻辑的平滑集成问题；
② 数据挖掘技术与特定数据存储类型的适应问题；
③ 大型数据的选择与规格化问题；
④ 数据挖掘系统的构架与交互式挖掘技术；
⑤ 数据挖掘语言与系统的可视化问题；
⑥ 数据挖掘理论与算法研究。

10. 按你对数据挖掘技术的了解，你认为它的研究将面临的主要挑战和对策是什么？

参考答案：略。

11. 你认为应该如何来理解 KDD 与 Data Mining 的关系？说明你的理由。

参考答案：关于 KDD 与 Data Mining 的关系有以下几种说法。

(1) KDD 看成数据挖掘的一个特例。这是早期比较流行的观点，在许多文献可以看到这种说法。因此，从这个意义上说，数据挖掘就是从数据库、数据仓库以及其他数据存储方式中挖掘有用知识的过程。这种描述强调了数据挖掘在源数据形式上的多样性。

(2) 数据挖掘是 KDD 过程的一个步骤(从狭义角度考虑)。这种观点得到大多数学者认同，有它的合理性。KDD 是一个广义的范畴，它包括数据清洗、数据集成、数据选择、数据转换、数据挖掘、模式生成及评估等一系列步骤。这样，可以把 KDD 看作是一些基本功能构件的系统化协同工作系统，而数据挖掘则是这个系统中的一个关键的部分。

(3) KDD 与 Data Mining 含义相同(从广义角度考虑)。有些人认为，KDD 与 Data Mining 只是叫法不一样，它们的含义基本相同。事实上，在现今文献的许多地方，这两个术语仍然不加区分地使用着。

从上面的描述中可以看出，数据挖掘概念可以在不同的技术层面上来理解，但是其核心仍然是从数据中挖掘知识。数据挖掘定义有广义和狭义之分。从广义的观点上，数据挖掘是从大型数据集中，挖掘隐含在其中的、人们事先不知道的、对决策有用的知识的过程。从狭义的观点上，可以定义数据挖掘是从特定形式的数据集中提炼知识的过程。

12. 解释将 Data Mining 理解为 KDD 整个过程的一个关键步骤地合理性。

参考答案：略。

13. 根据挖掘数据的对象不同，可以将数据挖掘技术进行分类，简述这些分类类型。

参考答案：根据挖掘数据的对象不同，数据挖掘技术可以分为关系型数据库挖掘、面向对象数据库挖掘、空间数据库挖掘、时态数据库挖掘、文本数据库挖掘、多媒体数据库挖掘、异质数据库挖掘、遗产数据库挖掘、Web 数据挖掘等。

14. 根据数据挖掘技术所依赖的主要技术来划分，数据挖掘技术有哪些主要的分类？简述这些类型的主要技术特点。

参考答案：略。

15. 粗糙集的知识形成主要是基于什么思想的？简述粗糙集理论中的信息系统、近似空间、下近似、上近似、约简等概念。

参考答案：粗糙集的知识形成思想可以概括为：一种类别对应于一个概念(类别一般表示为外延即集合，而概念常以如规则描述这样的内涵形式表示)，知识由概念组成；如果

某知识中含有不精确概念,则该知识不精确。粗糙集理论是一种研究不精确、不确定性知识的数学工具。

(1) 信息系统:一个信息系统 S 是一个四元组 $S=<U,A,V,f>$,其中 U 是对象(或事例)的有限集合,记为 $U=\{x_1,x_2,\cdots,x_n\}$;A 是属性的有限集合,记为 $A=\{A_1,A_2,\cdots,A_m\}$;V 是属性的值域集,记为 $V=\{V_1,V_2,\cdots,V_m\}$,其中 V_i 是属性 A_i 的值域;f 是信息函数(Information Function),即 $f:U\times A\rightarrow V$,$f(x_i,A_j)\in V_j$。

(2) 近似空间:近似空间有一个二元组 $<U,R(B)>$ 给出,其中 U 是对象(或事例)的有限集合,记为 $U=\{x_1,x_2,\cdots,x_n\}$;B 是 A 的属性子集,$R(B)$ 是 U 上的二元等价关系,即 $R(B)=\{(x_1,x_2)\mid f(x_1,b)=f(x_2,b),b\in B\}$。

(3) 下近似和上近似:对任意一个概念(或集合)O,B 是 U 的一个子集,O 的下近似定义为 $\underline{B}O=\{x\in U\mid[x]_{R(B)}\subset O\}$,其中 $[x]_{R(B)}$ 表示 x 在 $R(B)$ 上的等价类。O 的上近似定义为 $\overline{B}O=\{x\in U\mid[x]_{R(B)}\cap O\neq\varnothing\}$。一个概念(或集合)的下近似中的元素肯定属于该概念(或集合);而一个概念(或集合)的上近似概念(或集合)只是可能属于该概念。

(4) 约简:即极小属性集,也就是去掉约简中的任何一个属性,都将使得该属性集对应的规则覆盖反例,即导致规则与例子的不一致。

16. 简述粗糙集知识形成主要过程,为什么说它和数据挖掘技术在解决问题空间上有很大的重合性。

参考答案:略。

第 2 章　知识发现过程与应用结构

1. KDD 是一个多步骤的处理过程，它一般包含哪些基本阶段？简述各阶段的功能。

参考答案：KDD 是一个多步骤的处理过程，一般分为问题定义、数据抽取、数据预处理、数据挖掘以及模式评估等基本阶段。

(1) 问题定义阶段的功能：和领域专家以及最终用户紧密协作，一方面了解相关领域的有关情况，熟悉背景知识，弄清用户要求，确定挖掘的目标等要求；另一方面通过对各种学习算法的对比进而确定可用的学习算法。

(2) 数据抽取阶段的功能：选取相应的源数据库，并根据要求从数据库中提取相关的数据。

(3) 数据预处理阶段的功能：对前一阶段抽取的数据进行再加工，检查数据的完整性及数据的一致性。

(4) 数据挖掘阶段的功能：运用选定的数据挖掘算法，从数据中提取出用户所需要的知识。

(5) 模式评估阶段的功能：将 KDD 系统发现的知识以用户能了解的方式呈现，并且根据需要进行知识评价。如果发现知识和用户挖掘目标不一致，则重复以上阶段以最终获得可用的知识。

2. 为什么一个完整的知识发现要多种技术结合、多阶段集成。

参考答案：略

3. 简述在数据挖掘前要进行数据预处理的理由及其解决的主要问题。

参考答案：数据预处理包括：数据清洗、数据变换和数据归约等，是进行数据分析和挖掘的基础。如果所集成的数据不正确，数据挖掘算法输出的结果也必然不正确，这样形成的决策支持是不可靠的。因此，要提高挖掘结果的准确率，数据预处理是不可忽视的一步。

对数据进行预处理，一般需要对源数据进行再加工，检查数据的完整性及数据的一致性，对其中的噪音数据进行平滑，对丢失的数据进行填补，消除“脏”数据，消除重复记录等。

4. 为什么在知识发现过程中，要强调和用户交互的必要性？通常需要那些专长的技术人员支持？

参考答案：略

5. 阶梯处理过程模型是知识发现的基本模型，画出它的基本处理流程，并简要说明各阶段的任务。

参考答案：阶梯处理过程模型的基本处理流程如图 2-1 所示。

各阶段的主要任务是：

(1) 数据准备：了解相关领域的情况，弄清楚用户的要求，确定挖掘的总体目标和方法，并对原数据结构加以分析、确定数据选择原则等工作。

(2) 数据选择：从数据库中提取与 KDD 目标相关的数据。

(3) 数据预处理：主要是对上一阶段产生的数据进行再加工，检查数据的完整性及数据的一致性，对其中的噪音数据进行处理，对丢失的数据可以利用统计方法进行填补。对一

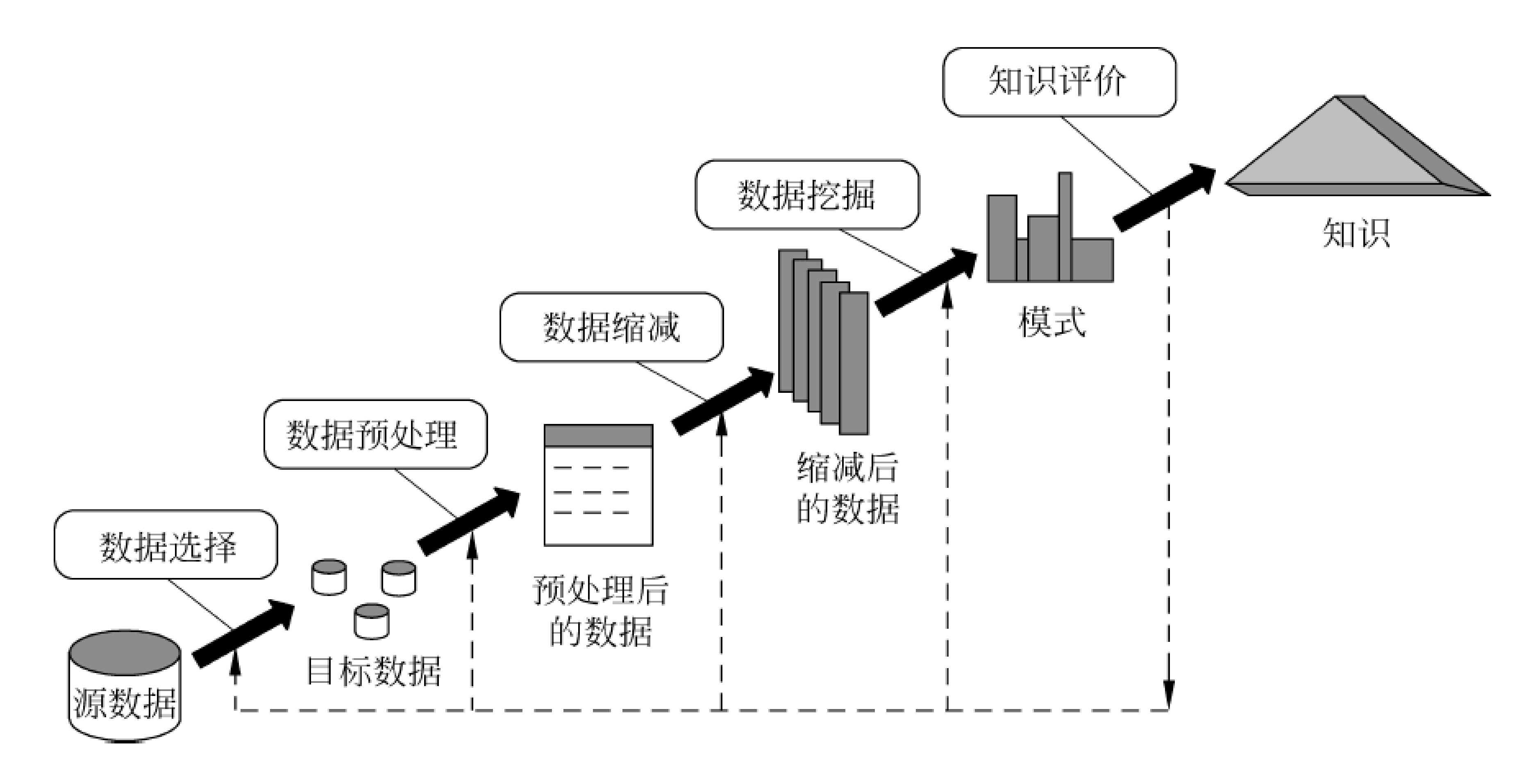


图 2-1　KDD 阶梯处理过程模型

些不适合于操作的数据进行必要的处理等。

(4) 数据缩减：对经过预处理的数据，根据知识发现的任务对数据进行抽取处理，使数据再次精简取其精华，更好地集中于用户挖掘目标上。

(5) 确定 KDD 的目标：根据数据挖掘的目标和用户的要求，确定 KDD 所发现的具体知识模式和类型(如分类、聚类、关联规则等)。

(6) 确定数据挖掘算法：根据上一阶段所确定的模式，选择合适的数据挖掘算法。(包括选取合适的参数、知识表示方式，并保证数据挖掘算法与整个 KDD 的评判标准相一致)。

(7) 数据挖掘：运用选定的算法，从数据中提取出用户所需要的知识。

(8) 模式解释：对发现的模式进行解释。在此过程中，为了取得更为有效的知识，可能会返回前面处理步骤中的某些步以改进结果，保证提取出的知识是有效和可用的。

(9) 知识评价：将发现的知识以用户能了解的方式呈现给用户。这期间也包含对知识的一致性的检查，以确信本次发现的知识不与以前发现的知识相抵触。

6. 简述螺旋处理过程模型相对于阶梯处理过程模型的优缺点。

参考答案：略

7. 简述以用户为中心的处理模型的基本思想。

参考答案：注重对用户与数据库交互的支持，用户根据数据库中的数据，提出一种假设模型，然后选择有关数据进行知识的挖掘，并不断对模型的数据进行调整优化，以提高数据挖掘的准确性和效率。因此，以用户为中心的处理模型的核心是将与用户的交互思想贯穿于数据挖掘的整个过程中。

8. 联机 KDD 模型需要解决哪些主要问题?

参考答案：略

9. 知识发现软件或工具的发展经历哪三个主要阶段? 简述他们的主要特点。

参考答案：知识发现软件或工具的发展经历了独立的知识发现软件、横向的知识发现

工具和纵向的知识发现解决方案三个主要阶段。

（1）独立的知识发现软件：这类软件要求用户必须对具体的数据挖掘技术和算法有相当的了解，还要手工负责大量的数据预处理工作。

（2）横向的知识发现工具：这些集成软件属于通用辅助工具范畴，可以帮助用户快速完成知识发现的不同阶段处理工作。使用这些工具，用户可以在数据挖掘和知识发现专家的指导和参与下开发对应的应用，起到了加速应用研制的作用。

（3）纵向的知识发现解决方案：这种方法的核心是针对特定的商业领域和商业逻辑提供完整的数据挖掘和知识发现解决方案。

10. 横向的知识发现工具集和纵向的知识发现解决方案的主要区别是什么？

参考答案：略。

11. 什么是知识发现项目的过程化管理？它的意义如何？

参考答案：知识发现是一个包括数据抽取、数据选择、数据挖掘以及模式评估等在内的系统化挖掘知识的过程。由于数据挖掘项目规模庞大，进行过程管理可以使其更加规范化。有效过程化管理是把实际问题分为若干子任务，在上一过程没有完成的情况下，下面的过程不能进行，以保证各个阶段的有序执行。

通过这样的模块化的管理过程，可以更好地完成数据挖掘任务，提高数据挖掘的效率和精度。

12. 简述强度挖掘的 I-MIN 过程模型的主要阶段和任务。

参考答案：略。

13. 简述数据挖掘语言的三种基本类型和特点。

参考答案：根据功能和侧重点不同，数据挖掘语言可以分为三种类型：数据挖掘查询语言、数据挖掘建模语言、通用数据挖掘语言。

（1）数据挖掘查询语言：遵循类似 SQL 的语法，通过数据挖掘的任务、功能以及其他约束的指定、知识形成和展示等系列工作，以类似于查询的形式输入到数据挖掘系统中，通过数据挖掘系统产生对应的结果。

（2）数据挖掘建模语言：是对数据挖掘模型进行描述和定义的语言。数据挖掘系统在模型定义和描述方面有标准可以遵循，那么各系统之间可以共享模型，既可以解决目前各数据挖掘系统之间封闭性的问题，又可以在其他应用系统中间嵌入数据挖掘模型，解决统一的知识发现描述问题。

（3）通用数据挖掘语言：通用数据挖掘语言合并了上述两种语言的特点，既具有定义模型的功能，又能作为查询语言与数据挖掘系统通信，进行交互式挖掘。

14. 为什么说数据挖掘语言研制对数据挖掘技术的发展是至关重要的？

参考答案：略。

第3章　关联规则挖掘理论和算法

1. 简单地描述下列英文缩写或短语的含义。

(1) Parallel Association Rule Mining

(2) Quantities Association Rule Mining

(3) Frequent Itemset

(4) Maximal Frequent Itemset

(5) Closed Itemset

参考答案:(1) 并行关联规则挖掘。是指利用并行处理技术、使用并行挖掘算法或在并行计算的环境下完成数据的高效挖掘工作。

(2) 数量关联规则挖掘。是指对含有诸如工资、价钱等非离散的数值属性的数据进行挖掘的技术。数量关联规则挖掘需要解决连续属性的离散化等问题,有更广泛的商业应用。

(3) 频繁项目集。是指出现频率高的项目对应的集合,反映交易数据中项目出现的频度信息。挖掘频繁项目集是关联规则挖掘的基础,许多关联规则挖掘方法是基于频繁项目集发现的。

(4) 最大频繁项目集。是指在频繁项目集中不出现相互包含的项目子集。最大频繁项目集可以使用最少的信息来保证频度信息的不丢失。

(5) 关闭(或闭和)项目集。简单地说,对于一个关闭项目集的任何元素,要么不被任何元素所包含,要么只被小于它的支持度的元素所包含。

2. 解释下列概念

(1) 多层次关联规则

(2) 多维关联规则

(3) 事务数据库

(4) 购物篮分析

(5) 强关联规则

参考答案:略。

3. 给出一个项目集 I_1 在数据集 D 上的支持度(Support)的定义,并直观地解释它的含义。

参考答案:设 $I_1 \subseteq I$,项目集 I_1 在数据集 D 上的支持度是包含 I_1 的事务在 D 中所占的百分比。直观上说,一个项目集在一个数据集 D 上的支持度反映了这个项目集在数据集中出现的频率。

4. 从统计学的观点说明一个项目集 I_1 在数据集 D 上的支持度的含义。

参考答案:略。

5. 满足什么样条件的项目集是频繁项目集和最大频繁项目集?

参考答案:对项目集 I 和事务数据库 D,D 中的所有大于或者等于满足用户指定的最小支持度的项目集称为频繁项目集。在最大频繁项目集,任何元素是频繁的而且不被其他元素所包含。

6. 以购物篮应用为例说明挖掘频繁项目集所蕴含的商业价值。

参考答案：略。

7. 给出一个规则的可信度(Confidence)的定义,并直观地解释它的含义。

参考答案：给定一个被讨论的项目集 I 和数据库 D,规则 $I_1 \Rightarrow I_2$ 的可信度是指包含 I_1 和 I_2 的事务数在只包含 I_1 的事务数所占的百分比。利用支持度定义可以描述为：

$$\text{Confidence}(I_1 \Rightarrow I_2) = \text{support}(I_1 \cup I_2)/\text{support}(I_1),$$

其中 $I_1, I_2 \subseteq I, I_1 \cap I_2 = \Phi$。

8. 以购物篮应用为例说明关联规则挖掘所蕴含的商业价值。

参考答案：略。

9. 一般地,在一个事务数据库中挖掘关联规则通过哪两个主要步骤完成,各步骤的主要任务和目标是什么?

参考答案：(1) 发现频繁项目集：通过用户给定的最小支持度,寻找所有频繁项目集,即满足 support 不小于最小支持度的所有项目子集。

(2) 生成关联规则：通过用户给定的最小可信度,在已经发现的最大频繁项目集中,寻找可信度不小于用户给定的最小可信度的关联规则。

10. 思考为什么事务数据库中挖掘关联规则一般要使用两个基本步骤?

参考答案：略。

11. 证明著名的 Agrawal 挖掘原理之一：频繁项目集的子集是频繁项目集。

参考答案：略。

证明：设 X 是一个项目集,事务数据库 T 中支持 X 的元组数为 s。对 X 的任一非空子集为 Y,设 T 中支持 Y 的元组数为 $s1$。

根据项目集支持度的定义,很容易知道：支持 X 的元组一定支持 Y,所以 $s1 \geqslant s$,即

$$\text{support}(Y) \geqslant \text{support}(X)。$$

按假设,项目集 X 是频繁项目集,即

$$\text{support}(X) \geqslant \text{minsupport},$$

所以 $\text{support}(Y) \geqslant \text{support}(X) \geqslant \text{minsupport}$,因此 Y 是频繁项目集。

12. 证明著名的 Agrawal 挖掘原理之一：非频繁项目集的超集是非频繁项目集。

参考答案：略。

13. 给定如表 3-1 所示的一个事务数据库,写出 Apriori 算法生成频繁项目集的过程(假设 Minsuport=50%)。

表 3-1　事务数据库示例 1

TID	Itemset	TID	Itemset
1	a,c,d,e,f	4	a,c,d,e
2	b,c,f	5	a,b,d,e,f
3	a,d,f		

参考答案:

$L1$ 生成:$C1=\{(a,4)(b,2)(c,3)(d,4)(e,3)(f,4)\}$;$L1=\{a,c,d,e,f\}$。

$L2$ 生成:$C2=\{(ac,2)(ad,4)(ae,3)(af,3)(cd,2)(ce,2)(cf,2)(de,3)(df,3)(ef,2)\}$;$L2=\{ad,ae,af,de,df\}$。

$L3$ 生成:$C3=\{(ade,3)(adf,3)(def,2)\}$;$L3=\{ade,adf\}$。

$L4$ 生成:$C4$:$\{(adef,2)\}$;$L4=\varnothing$。

$L5$ 生成:$C5=\varnothing,L5=\varnothing$。

结束后,最大频繁项目集为$\{ade,adf\}$。

14. 给定如表 3-2 所示的一个事务数据库,写出 Apriori 算法生成频繁项目集的过程(假设 Minsuport=40%)。

表 3-2 事务数据库示例 2

TID	Itemset	TID	Itemset
1	1,3,4	4	2,5
2	2,3,4,5	5	1,2,4,6,7
3	1,3,5,7	6	2,4,6

参考答案:略。

15. 对上面的第 13 题所生成的最大频繁项目集,跟踪 Rule-generate 来生成对应的关联规则(设 minconfidence=80%)。

参考答案:生成过程如表 3-3 所示。

表 3-3 生成过程

序号	l_k	x_{m-1}	confidence	support	规则(是否是强规则)
1	ade	ad	75%	60%	$ad \to e$ 否
2	ade	a	75%	60%	$a \to de$ 否
3	ade	d	75%	60%	$d \to ae$ 否
4	ade	ae	100%	60%	$ae \to d$ 是
5	ade	e	100%	60%	$e \to ad$ 是
6	ade	de	100%	60%	$de \to a$ 是
7	adf	ad	75%	60%	$ad \to f$ 否
8	adf	a	75%	60%	$a \to df$ 否
9	adf	d	75%	60%	$d \to af$ 否
10	adf	af	100%	60%	$af \to d$ 是
11	adf	f	75%	60%	$f \to ad$ 否
12	adf	df	100%	60%	$df \to a$ 是

16. 对上面的第 14 题所生成的最大频繁项目集,跟踪 Rule-generate 来生成对应的关联规则(设 minconfidence=60%)。

参考答案：略。

17. Apriori 算法的主要性能瓶颈是什么？

参考答案：Apriori 算法的主要性能瓶颈有：

(1) 多次扫描事务数据库，需要很大的 I/O 负载；

(2) 可能产生庞大的候选集，由 L_{k-1} 产生 k-候选集 C_k 是指数增长的。

18. 针对 Apriori 算法的主要性能瓶颈提出你的改进想法。

参考答案：略。

19. 基于数据分割(Partition)的方法可以改善 Apriori 算法的效率。阐述它的理由。

参考答案：(1) 合理利用主存空间。数据分割为块内数据一次性导入主存提供机会，因而提高对大容量数据集的挖掘效率。

(2) 支持并行挖掘算法。

20. 基于采样(Sampling)的方法可以改善 Apriori 算法的效率。阐述它的理由。

参考答案：略。

21. 基于散列(Hash)的方法，可以改善 Apriori 算法的效率。阐述它的理由。

参考答案：使用散列的方法产生频繁项目集，可以改善 Apriori 算法的效率，主要是因为散列拥有能够快速查找元素的特性。这种方法把扫描的项目放到不同的哈希桶中，每个项目集最多只可能在一个特定的桶中。这样可以对每个桶中的项目子集进行测试，减少了候选集生成的代价。

22. 除了上面提到的技术可以用于改善 Apriori 算法的效率以外，你认为还有那些技术可以被应用来解决这个问题。

参考答案：略。

23. 一个项目集是闭合的(Closed)，简单地讲它应该满足什么条件？

参考答案：一个项目集 C 是闭合的，当且仅当对于在 C 中的任何元素，不可能在 C 中存在小于或等于它的支持度的子集。

24. 为什么说在闭合项目集格空间里讨论关联规则挖掘问题要比 Apriori 算法效率高？

参考答案：略。

25. FP-tree 的算法是一个 2 次数据库扫描算法，这个算法的基本思想是什么？

参考答案：FP-tree 算法只进行 2 次数据库扫描。它不使用候选集，直接压缩数据库成一个频繁模式树，最后通过这棵树生成关联规则。

用 FP-tree 挖掘频繁集基本思想是分而治之，即用 FP-tree 递归增长形成频繁集。

26. 比较 Apriori 算法，阐述 FP-tree 的算法的优缺点。

参考答案：略。

27. 给定如表 3-4 所示的一个事务数据库，画出 FP-tree 树的生成过程。

参考答案：(1) 首先扫描数据库按照支持度将序排列生成索引，如表 3-5 所示。

表 3-4　事务数据库示例 3

TID	Itemset
1	*a*,*b*,*c*
2	*b*,*c*,*d*,*e*
3	*a*,*c*,*e*
4	*b*,*c*,*d*
5	*b*,*c*,*d*,*e*

表 3-5　索引表

Item	SCP
c	5
b	4
d	3
e	3
a	2

(2) 扫描数据库,对每个事务进行树的增长并改变支持度,其演化过程如图 3-1 所示。

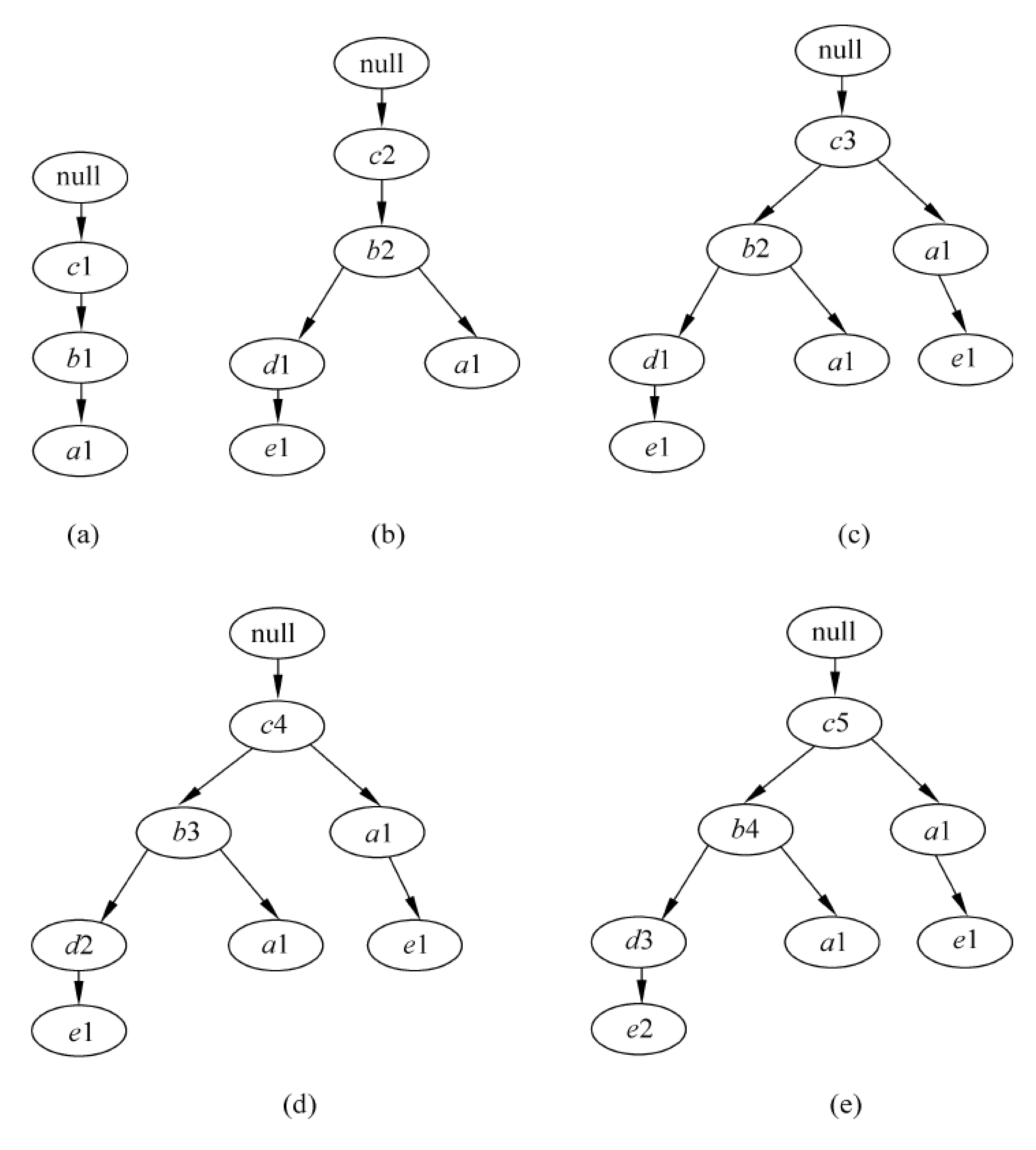


图 3-1　FP-tree 生成过程示意图

(3) 连接索引表,生成最终的结果,如图 3-2 所示。

28. 给定如表 3-6 所示的一个事务数据库,画出 FP-tree 树的生成过程。

表 3-6　事务数据库示例 4

TID	Itemset	TID	Itemset
1	*B*,*C*,*D*,*E*	4	*C*,*D*,*E*,*F*
2	*A*,*C*,*E*	5	*A*,*B*,*C*,*D*,*E*,*F*
3	*A*,*B*,*C*,*E*		

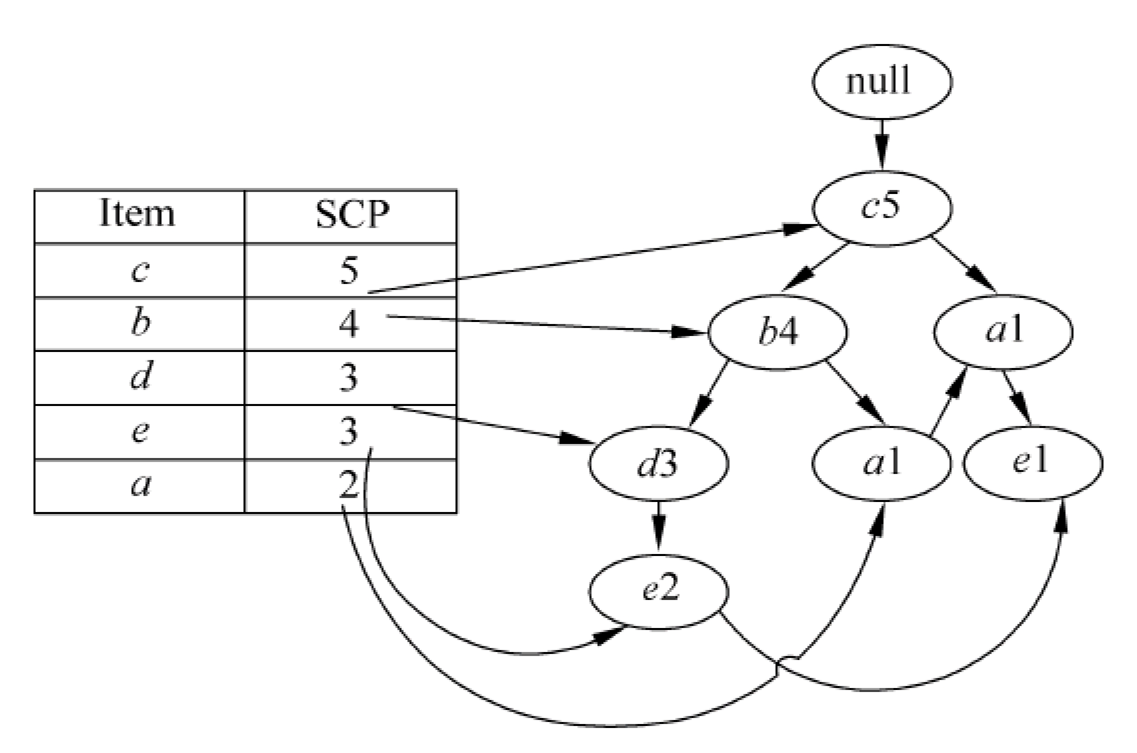


图 3-2 FP-tree 示意图

参考答案：略。

29. 衡量关联规则挖掘结果的有效性应该从哪些方面加以考虑？简述其理由。

参考答案：(1) 准确性：挖掘出的规则必须反映数据的实际情况。尽管规则不可能是100%适用的，但是必须要在一定的可信度内。

(2) 实用性：挖掘出的规则必须是简洁可用的，而且是针对挖掘目标的。不能说有 100 条规则，其中 50 条与商业目标无关，30 条用户无法理解。

(3) 新颖性：挖掘出的关联规则可以为用户提供新的有价值信息。如果它们是用户事先就知道的，那么这样的规则即使再正确也是毫无价值的。

30. 为什么说用户从主观层面上为关联规则挖掘设定约束条件是必要的？应该从几个方面来考虑这个问题？

参考答案：略。

31. 简述约束在数据挖掘中的作用。

参考答案：归纳起来，约束在数据挖掘中的使用可以在如下方面起到关键作用。

(1) 聚焦挖掘任务，提高挖掘效率：利用约束，把具体的挖掘任务转换成对系统工作的控制，从而使挖掘工作按着期望的方向发展。通过人机交互和探索实验，可以快速聚焦挖掘任务，进而提高挖掘效率。

(2) 保证挖掘的精确性：约束的使用可以帮助发现问题，并及时加以调整，使知识发现的各个阶段按着正确的方向发展。

(3) 控制系统的使用规模：约束数据挖掘的思想为系统的增量式扩充提供条件。当基本的原则和目标确定后，可以把一些有待验证和优化的问题以约束参数的形式交互式输入，通过实验找到最佳值。在挖掘阶段，可以针对不同的子目标进行约束，快速聚焦问题，加快知识形成进程。

32. 从挖掘所使用约束的类型看，可以把用于关联规则挖掘的约束分为哪些类型？通过实例来理解这些类型的应用。

参考答案：略。

33. 多层次关联规则挖掘的有两种基本策略，简述它们可能存在的主要问题及相关对策。

参考答案：多层次关联规则挖掘有以下两种基本的设置支持度的策略。

(1) 统一的最小支持度：对于所有层次，都使用同一个最小支持度。这样对于用户和算法实现来说，相对容易，而且很容易支持层间的关联规则生成。但是弊端也是显然的。首先，不同层次可能考虑问题的精度不同、面向的用户群不同。对于一些用户，可能觉得支持度太小，产生了过多不感兴趣的规则。而对于另外的用户来说，又认为支持度太大，有用信息丢失过多。

(2) 不同层次使用不同的最小支持度：每个层次都有自己的最小支持度。较低层次的最小支持度相对较小，而较高层次的最小支持度相对较大。这种方法增加了挖掘的灵活性。但是，也留下了许多相关问题需要解决。首先，不同层次间的支持度应该有所关联，只有正确地刻画这种联系或找到转换方法，才能使生成的关联规则相对客观。另外，由于具有不同的支持度，层间的关联规则挖掘也是必须解决的问题。例如，有人提出层间关联规则应该根据较低层次的最小支持度来定。

34. 为什么多层次关联规则挖掘可能产生规则的冗余问题，你认为应该如何有效地避免这些冗余问题可能带来的副作用。

参考答案：略。

35. 举例说明单维关联规则和多维关联规则的区别。

参考答案：多维和单维关联规则的主要区别在于维数。比如，“年龄(X，20～30)? 职业(X，学生)＝＞购买(X，笔记本电脑)”。这里涉及三个维：年龄、职业、购买，所以它被称为多维关联规则。而又比如“啤酒＝＞尿布”这样的关联规则只涉及“购买”这一单一维，因此被称为单维关联规则。

36. 思考多维关联规则挖掘所带来的主要挑战。

参考答案：略。

37. 数量关联数规则要解决什么样的问题？简述处理数值属性的基本方法。

参考答案：数量关联规则挖掘有许多问题值得讨论。目前比较集中和急需解决的关键问题有下面三个主要方面：

(1) 连续数值属性的处理；

(2) 规则的优化；

(3) 提高挖掘效率。

一般而言，连续数值属性的处理有两种基本的方法：

(1) 对数值属性进行离散化处理，这样就把连续的数值属性转变成布尔型属性，因此可以利用已有的方法和算法。这是目前研究比较多的方法。比较著名的有等深度桶方法、部分 K 度完全方法等。

(2) 不直接对数值属性离散化，而是采用统计或模糊方法直接处理它们。直接用数值字段中的原始数据进行分析，可能结合多层次关联规则的概念，在多个层次之间进行比较从而得出一些有用的规则。

38. 简述数量关联规则挖掘的一般步骤。

参考答案：略。

第4章　分类方法

1. 简单地描述下列英文缩写或短语的含义。

(1) Data Classification

(2) k-Nearest Neighbors

(3) Decision Tree

(4) Entropy

(5) Posterior Probability

参考答案：(1) 数据分类。用分类模型(也常常称作分类器)把数据库中的数据项映射到给定类别中的某一个类别。

(2) k-最临近方法。它是一种基于距离的分类算法。

(3) 决策树。决策树是一个类似于流程图的树结构，其中每个内部结点表示在一个属性上的测试，每个分支代表一个测试输出，而每个树叶结点代表类或类分布。树的最顶层结点是根结点。决策树表示方法是分类中应用最广泛的方法之一。

(4) 熵。在信息论中，熵是一种信息度量单位。在决策树构造算法中根据熵值来计算信息增益。

(5) 后验概率。后验概率又被称为条件概率，是在已知结果发生的情况下，求导致结果的某种原因的可能性的大小。比如求 $P(H|X)$，当 $P(H)$、$P(X)$、$P(X|H)$ 容易求得时，可以由贝叶斯公式得出 $P(H|X)=\frac{P(X|H)P(H)}{P(X)}$，这里 $P(H)$ 是先验概率(Prior Probability)，$P(X|H)$ 表示假设 H 成立的情况下观察到 X 的概率，$P(H|X)$ 是后验概率(或称条件 X 下 H 的后验概率)。

2. 简述数据分类的概念。

参考答案：略。

3. 数据分类分为哪两个步骤？简述每步的基本任务。

参考答案：分类归结为模型建立和使用模型进行分类两个步骤。

第一步的基本任务是建立一个模型并描述预定的数据类集；第二步的基本任务是评估模型的预测准确率，用准确率可以接受的模型对类标号未知的数据进行分类。

4. 简述基于距离的分类算法的主要思想。

参考答案：略。

5. 简述 k-最临近方法的主要思想。

参考答案：计算每个训练数据(每个训练数据都有一个唯一的类别标识)到待分类元组的距离，取和待分类元组距离最近的 k 个训练数据，k 个数据中哪个类别的训练数据占多数，则待分类元组就属于哪个类别。

6. 简述决策树算法的主要步骤。

参考答案：略。

7. 决策树容易转换成分类规则,试把图 4-1 所示的决策树转换成分类规则(假定决策属性为 buys_computer)。

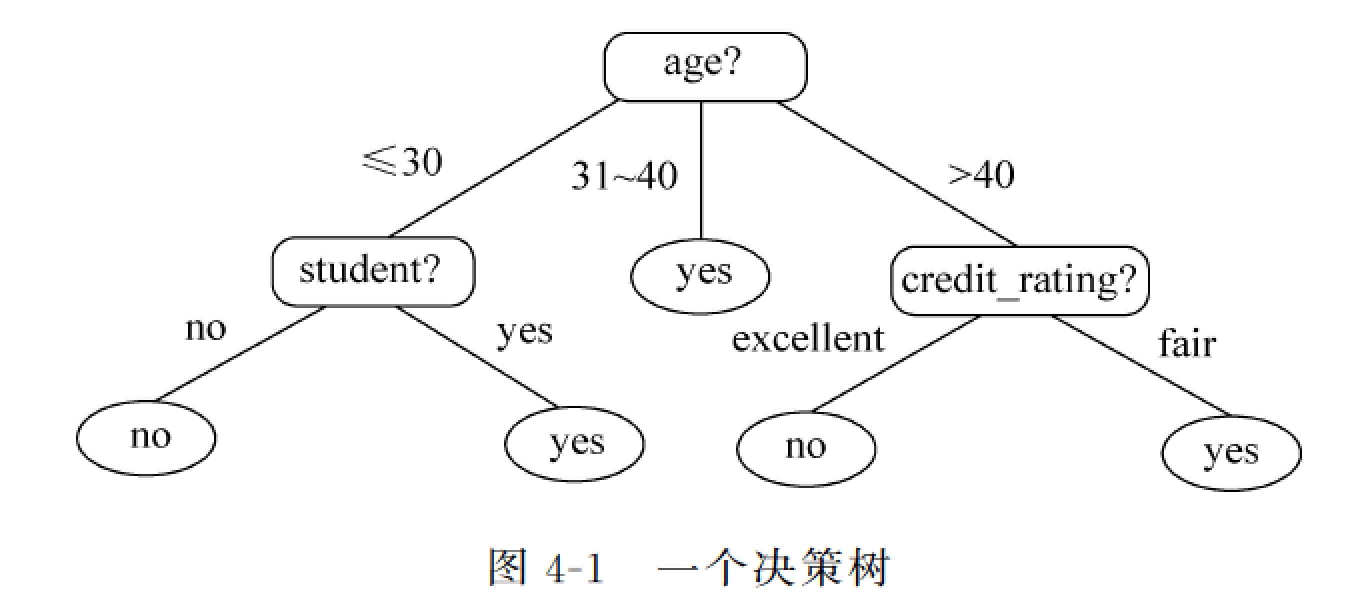


图 4-1 一个决策树

参考答案:

If age≤30 and student=no Then buys_computer=no

If age≤30 and student=yes Then buys_cpmputer=yes

If age>30 and age≤40 Then buys_computer=yes

If age>40 and credit_rating=excellent Then buys_computer=no

If age>40 and credit_rating=fair Then buys_computer=yes

8. 在决策树算法中,剪枝的作用是什么?

参考答案:略。

9. 表 4-1 给出了一个关于配眼镜的一个决策分类所需要的数据。数据集包含以下 5 个属性:

- age:{young,pre-presbyopic,presbyopic}。
- astigmatism:{no,yes}。
- spectacle-prescrip:{myope,hypermetrope}。
- tear-prod-rate:{reduced,normal}。
- contact-lenses:{soft,none,hard}。

contact-lenses 是决策属性,手动模拟 ID3 算法来实现决策过程。

表 4-1 训练数据集

	age	spectacle-prescrip	astigmatism	tear-prod-rate	contact-lenses
1	young	myope	no	reduced	none
2	young	myope	no	normal	soft
3	young	myope	yes	reduced	none
4	young	myope	yes	normal	hard
5	young	hypermetrope	no	reduced	none
6	young	hypermetrope	no	normal	soft
7	young	hypermetrope	yes	reduced	none
8	young	hypermetrope	yes	normal	hard
9	pre-presbyopic	myope	no	reduced	none
10	pre-presbyopic	myope	no	normal	soft

续表

	age	spectacle-prescrip	astigmatism	tear-prod-rate	contact-lenses
11	pre-presbyopic	myope	yes	reduced	none
12	pre-presbyopic	myope	yes	normal	hard
13	pre-presbyopic	hypermetrope	no	reduced	none
14	pre-presbyopic	hypermetrope	no	normal	soft
15	pre-presbyopic	hypermetrope	yes	reduced	none
16	pre-presbyopic	hypermetrope	yes	normal	none
17	presbyopic	myope	no	reduced	none
18	presbyopic	myope	no	normal	none
19	presbyopic	myope	yes	reduced	none
20	presbyopic	myope	yes	normal	hard
21	presbyopic	hypermetrope	no	reduced	none
22	presbyopic	hypermetrope	no	normal	soft
23	presbyopic	hypermetrope	yes	reduced	none
24	presbyopic	hypermetrope	yes	normal	none

参考答案：(1) 计算给定样本 contact-lenses 分类所需的期望信息。

最终需要分类的属性为 contact-lenses，它有 3 个不同取值 none、soft 和 hard，none 有 15 个样本，soft 有 5 个样本，hard 有 4 个样本。因此，给定样本 contact-lenses 分类所需的期望信息：

$$I(s_1,s_2,s_3)=I(15,5,4)=-\frac{15}{24}\log_2\frac{15}{24}-\frac{5}{24}\log_2\frac{5}{24}-\frac{4}{24}\log_2\frac{4}{24}$$
$$=0.424+0.471+0.431=1.326。$$

(2) 计算每个属性的熵。

观察 age 的每个样本值 young、pre-presbyopic、presbyopic 的分布，具体情况如表 4-2 所示。

表 4-2　age 的样本值分布

contact-lenses / age	none	soft	hard
young	4	2	2
pre-presbyopic	5	2	1
presbyopic	6	1	1

对于 age=young，$s_{11}=4,s_{21}=2,s_{31}=2$，

$$I(s_{11},s_{21},s_{31})=I(4,2,2)=-\frac{4}{8}\log_2\frac{4}{8}-\frac{2}{8}\log_2\frac{2}{8}-\frac{2}{8}\log_2\frac{2}{8}$$
$$=0.5+0.5+0.5=1.5;$$

对于 age=pre-presbyopic，$s_{12}=5,s_{22}=2,s_{32}=1$，

$$I(s_{12},s_{22},s_{32})=I(5,2,1)=-\frac{5}{8}\log_2\frac{5}{8}-\frac{2}{8}\log_2\frac{2}{8}-\frac{1}{8}\log_2\frac{1}{8}$$
$$=0.424+0.5+0.375=1.299;$$

对于 age=presbyopic，$s_{13}=6,s_{23}=1,s_{33}=1$，

$$I(s_{13},s_{23},s_{33})=I(6,1,1)=-\frac{6}{8}\log_2\frac{6}{8}-\frac{1}{8}\log_2\frac{1}{8}-\frac{1}{8}\log_2\frac{1}{8}$$
$$=0.311+0.375+0.375=1.061;$$

所以，如果样本按 age 划分，对一个给定的样本分类对应的熵为：

$$E(\text{age})=\frac{8}{24}I(s_{11},s_{21},s_{31})+\frac{8}{24}I(s_{12},s_{22},s_{32})+\frac{8}{24}I(s_{13},s_{23},s_{33})$$
$$=\frac{1}{3}(1.5+1.299+1.061)=1.287。$$

因此，如果样本按 age 划分，得到的信息增益是：

$$Gain(age)=I(s_1,s_2,s_3)-E(age)=1.326-1.287=0.039。$$

观察 spectacle-prescrip 的每个样本值 myope、hypermetrope 的分布，具体情况如表 4-3 所示。

表 4-3 spectacle-prescrip 的样本值分布

contact-lenses / spectacle-prescrip	none	soft	hard
myope	7	2	3
hypermetrope	8	3	1

对于 spectacle-prescrip=myope，$s_{11}=7$，$s_{21}=2$，$s_{31}=3$，

$$I(s_{11},s_{21},s_{31})=I(7,2,3)=-\frac{7}{12}\log_2\frac{7}{12}-\frac{2}{12}\log_2\frac{2}{12}-\frac{3}{12}\log_2\frac{3}{12}$$
$$=0.454+0.431+0.5=1.385;$$

对于 spectacle-prescrip=hypermetrope，$s_{12}=8$，$_{22}=3$，$s_{32}=1$，

$$I(s_{12},s_{22},s_{32})=I(8,3,1)=-\frac{8}{12}\log_2\frac{8}{12}-\frac{3}{12}\log_2\frac{3}{12}-\frac{1}{12}\log_2\frac{1}{12}$$
$$=0.39+0.5+0.299=1.189;$$

所以，如果样本按 spectacle-prescrip 划分，对一个给定的样本分类对应的熵为：

$$E(\text{spectacle-prescrip})=\frac{12}{24}I(s_{11},s_{21},s_{31})+\frac{12}{24}I(s_{12},s_{22},s_{32})=\frac{1}{2}(1.385+1.189)=1.287。$$

因此，假如按 spectacle-prescrip 划分，信息增益是：

$$Gain(\text{spectacle-prescrip})=I(s_1,s_2,s_3)-E(\text{spectacle-prescrip})$$
$$=1.326-1.287=0.039。$$

观察 astigmatism 的每个样本值 no、yes 的分布，具体情况如表 4-4 所示。

表 4-4 astigmatism 的样本值分布

contact-lenses / astigmatism	none	soft	hard
no	7	5	0
yes	8	0	4

对于 astigmatism=no，$s_{11}=7$，$s_{21}=5$，$s_{31}=0$，

$$I(s_{11},s_{21},s_{31})=I(7,5,0)=-\frac{7}{12}\log_2\frac{7}{12}-\frac{5}{12}\log_2\frac{5}{12}-\frac{0}{12}\log_2\frac{0}{12}$$

$$= 0.454 + 0.53 + 0 = 0.984;$$

对于 astigmatism＝yes，$s_{12}=8$，$s_{22}=0$，$s_{32}=4$，

$$I(s_{12},s_{22},s_{32}) = I(8,3,1) = -\frac{8}{12}\log_2\frac{8}{12} - \frac{0}{12}\log_2\frac{0}{12} - \frac{4}{12}\log_2\frac{4}{12}$$
$$= 0.39 + 0 + 0.53 = 0.92;$$

所以，如果样本按 astigmatism 划分，对一个给定的样本分类对应的熵为：

$$E(\text{astigmatism}) = \frac{12}{24}I(s_{11},s_{21},s_{31}) + \frac{12}{24}I(s_{12},s_{22},s_{32}) = \frac{1}{2}(0.984 + 0.92) = 0.952。$$

因此，假如按 astigmatism 划分，信息增益是：

$$Gain(\text{astigmatism}) = I(s_1,s_2,s_3) - E(\text{astigmatism}) = 1.326 - 0.952 = 0.374。$$

观察 tear-prod-rate 的每个样本值 reduced、normal 的分布，具体情况如表 4-5 所示。

表 4-5　tear-prod-rate 的样本值分布

tear-prod-rate \ contact-lenses	none	soft	hard
reduced	12	0	0
normal	3	5	4

对于 tear-prod-rate＝reduced，$s_{11}=12$，$s_{21}=0$，$s_{31}=0$，

$$I(s_{11},s_{21},s_{31}) = I(12,0,0) = -\frac{12}{12}\log_2\frac{12}{12} - \frac{0}{12}\log_2\frac{0}{12} - \frac{0}{12}\log_2\frac{0}{12} = 0 + 0 + 0 = 0;$$

对于 tear-prod-rate＝normal，$s_{12}=3$，$s_{22}=5$，$s_{32}=4$，

$$I(s_{12},s_{22},s_{32}) = I(3,5,4) = -\frac{3}{12}\log_2\frac{3}{12} - \frac{5}{12}\log_2\frac{5}{12} - \frac{4}{12}\log_2\frac{4}{12}$$
$$= 0.5 + 0.53 + 0.53 = 1.56。$$

所以，如果样本按 tear-prod-rate 划分，对一个给定的样本分类对应的熵为：

$$E(\text{tear-prod-rate}) = \frac{12}{24}I(s_{11},s_{21},s_{31}) + \frac{12}{24}I(s_{12},s_{22},s_{32}) = \frac{1}{2}(0 + 1.56) = 0.78。$$

因此，假如按 tear-prod-rate 划分，信息增益是：

$$Gain(\text{tear-prod-rate}) = I(s_1,s_2,s_3) - E(\text{tear-prod-rate}) = 1.326 - 0.78 = 0.546。$$

由于 tear-prod-rate 在属性中具有最高的信息增益，所以它首先被选作测试属性，以此创建一个结点，用 tear-prod-rate 标记，并对于每个属性值，引出一个分支，如图 4-2 所示。

(3) 进一步生成左子树和右子树。

对于 tear-prod-rate＝reduced 的所有元组，其类别标记均为 none。所以，根据决策树生成算法步骤 2 和步骤 3，得到一个叶子结点，类别标记为 contact-lenses＝none。

对于 tear-prod-rate＝normal 的右子树中的所有元组，首先计算出给定样本 contact-lenses 分类所需的期望信息：

$$I(s_1,s_2,s_3) = I(3,5,4) = -\frac{3}{12}\log_2\frac{3}{12} - \frac{5}{12}\log_2\frac{5}{12} - \frac{4}{12}\log_2\frac{4}{12}$$
$$= 0.5 + 0.53 + 0.53 = 1.56。$$

对于 tear-prod-rate＝normal 的右子树中的所有元组（对应图 4-2 中的 T_2），计算其他三个属性的信息增益。

tear-prod-rate

=reduced　　　　　　　　= normal

T_1

age	spectacle-prescrip	astigmatism	contact-lenses
young	myope	no	none
young	myope	yes	none
young	hypermetrope	no	none
young	hypermetrope	yes	none
pre-presbyopic	rmyope	no	none
pre-presbyopic	rmyope	yes	none
pre-presbyopic	hypermetrope	no	none
pre-presbyopic	hypermetrope	yes	none
presboyopic	rmyope	no	none
presboyopic	rmyope	yes	none
presboyopic	hypermetrope	no	none
presboyopic	hypermetrope	yes	none

T_2

age	spectacle-prescrip	astigmatism	contact-lenses
young	myope	no	soft
young	myope	yes	hard
young	hypermetrope	no	soft
young	hypermetrope	yes	hard
pre-presbyopic	rmyope	no	soft
pre-presbyopic	rmyope	yes	hard
pre-presbyopic	hypermetrope	no	soft
pre-presbyopic	hypermetrope	yes	none
presboyopic	rmyope	no	none
presboyopic	rmyope	yes	hard
presboyopic	hypermetrope	no	soft
presboyopic	hypermetrope	yes	none

图 4-2　tear-prod-rate 结点及其分支

观察 age 的每个样本值 young、pre-presbyopic、presbyopic 的分布，具体情况如表 4-6 所示。

表 4-6　age 的样本值分布

contact-lenses / age	none	soft	hard
young	0	2	2
pre-presbyopic	1	2	1
presbyopic	2	1	1

对于 age=young，$s_{11}=0, s_{21}=2, s_{31}=2$，

$$I(s_{11},s_{21},s_{31})=I(0,2,2)=-\frac{0}{4}\log_2\frac{0}{4}-\frac{2}{4}\log_2\frac{2}{4}-\frac{2}{4}\log_2\frac{2}{4}$$

$$=0+0.5+0.5=1;$$

对于 age=pre-presbyopic，$s_{12}=1, s_{22}=2, s_{32}=1$，

$$I(s_{12},s_{22},s_{32})=I(1,2,1)=-\frac{1}{4}\log_2\frac{1}{4}-\frac{2}{4}\log_2\frac{2}{4}-\frac{1}{4}\log_2\frac{1}{4}$$

$$=0.5+0.5+0.5=1.5;$$

对于 age=presbyopic，$s_{13}=2, s_{23}=1, s_{33}=1$，

$$I(s_{13},s_{23},s_{33})=I(2,1,1)=-\frac{2}{4}\log_2\frac{2}{4}-\frac{1}{4}\log_2\frac{1}{4}-\frac{1}{4}\log_2\frac{1}{4}$$

$$=0.5+0.5+0.5=1.5。$$

因此，如果样本按 age 划分，对一个给定的样本分类对应的熵和信息增益为：

$$E(\text{age}) = \frac{4}{12}I(s_{11},s_{21},s_{31}) + \frac{4}{12}I(s_{12},s_{22},s_{32}) + \frac{4}{12}I(s_{13},s_{23},s_{33})$$
$$= \frac{1}{3}(1+1.5+1.5) = 1.33;$$
$$Gain(age) = I(s_1,s_2,s_3) - E(age) = 1.56 - 1.33 = 0.23。$$

观察 spectacle-prescrip 的每个样本值 myope、hypermetrope 的分布(对应图 4-2 中的 T_2),具体情况如表 4-7 所示。

表 4-7　spectacle-prescrip 的样本值分布

spectacle-prescrip \ contact-lenses	none	soft	hard
myope	1	2	3
hypermetrope	2	3	1

对于 spectacle-prescrip=myope,$s_{11}=1$,$s_{21}=2$,$s_{31}=3$,

$$I(s_{11},s_{21},s_{31}) = I(1,2,3) = -\frac{1}{6}\log_2\frac{1}{6} - \frac{2}{6}\log_2\frac{2}{6} - \frac{3}{6}\log_2\frac{3}{6}$$
$$= 0.431 + 0.53 + 0.5 = 1.461;$$

对于 spectacle-prescrip=hypermetrope,$s_{12}=2$,$s_{22}=3$,$s_{32}=1$,

$$I(s_{12},s_{22},s_{32}) = I(2,3,1) = -\frac{2}{6}\log_2\frac{2}{6} - \frac{3}{6}\log_2\frac{3}{6} - \frac{1}{6}\log_2\frac{1}{6}$$
$$= 0.53 + 0.5 + 0.431 = 1.461。$$

因此,如果样本按 spectacle-prescrip 划分,对一个给定的样本分类对应的熵和信息增益为:

$$E(\text{spectacle-prescrip}) = \frac{6}{12}I(s_{11},s_{21},s_{31}) + \frac{6}{12}I(s_{12},s_{22},s_{32})$$
$$= \frac{1}{2}(1.461 + 1.461) = 1.461;$$
$$Gain(\text{spectacle-prescrip}) = I(s_1,s_2,s_3) - E(\text{spectacle-prescrip}) = 1.56 - 1.461 = 0.099。$$

观察 astigmatism 的每个样本值 no、yes 的分布(对应图 4-2 中的 T_2),具体情况如表 4-8 所示。

表 4-8　astigmatism 的样本值分布

astigmatism \ contact-lenses	none	soft	hard
no	1	5	0
yes	2	0	4

对于 astigmatism=no,$s_{11}=1$,$s_{21}=5$,$s_{31}=0$,

$$I(s_{11},s_{21},s_{31}) = I(1,5,0) = -\frac{1}{6}\log_2\frac{1}{6} - \frac{5}{6}\log_2\frac{5}{6} - \frac{0}{6}\log_2\frac{0}{6}$$
$$= 0.431 + 0.219 + 0 = 0.65;$$

对于 astigmatism=no,$s_{12}=2$,$s_{22}=0$,$s_{32}=4$,

$$I(s_{12},s_{22},s_{32}) = I(8,3,1) = -\frac{2}{6}\log_2\frac{2}{6} - \frac{0}{6}\log_2\frac{0}{6} - \frac{4}{6}\log_2\frac{4}{6}$$
$$= 0.53 + 0 + 0.39 = 0.92。$$

因此,如果样本按 astigmatism 划分,对一个给定的样本分类对应的熵和信息增益为:

$$E(\text{astigmatism}) = \frac{6}{12}I(s_{11},s_{21},s_{31}) + \frac{6}{12}I(s_{12},s_{22},s_{32}) = \frac{1}{2}(0.65+0.92) = 0.785;$$

$$Gain(\text{astigmatism}) = I(s_1,s_2,s_3) - E(\text{astigmatism}) = 1.56 - 0.785 = 0.775。$$

由于 astigmatism 在属性中具有最高的信息增益,所以它被选作测试属性。并以此创建一个结点,用 astigmatism 标记,并对于每个属性值,引出一个分支,数据集被划分成两个子集。图 4-3 给出了 astigmatism 结点及其分支。

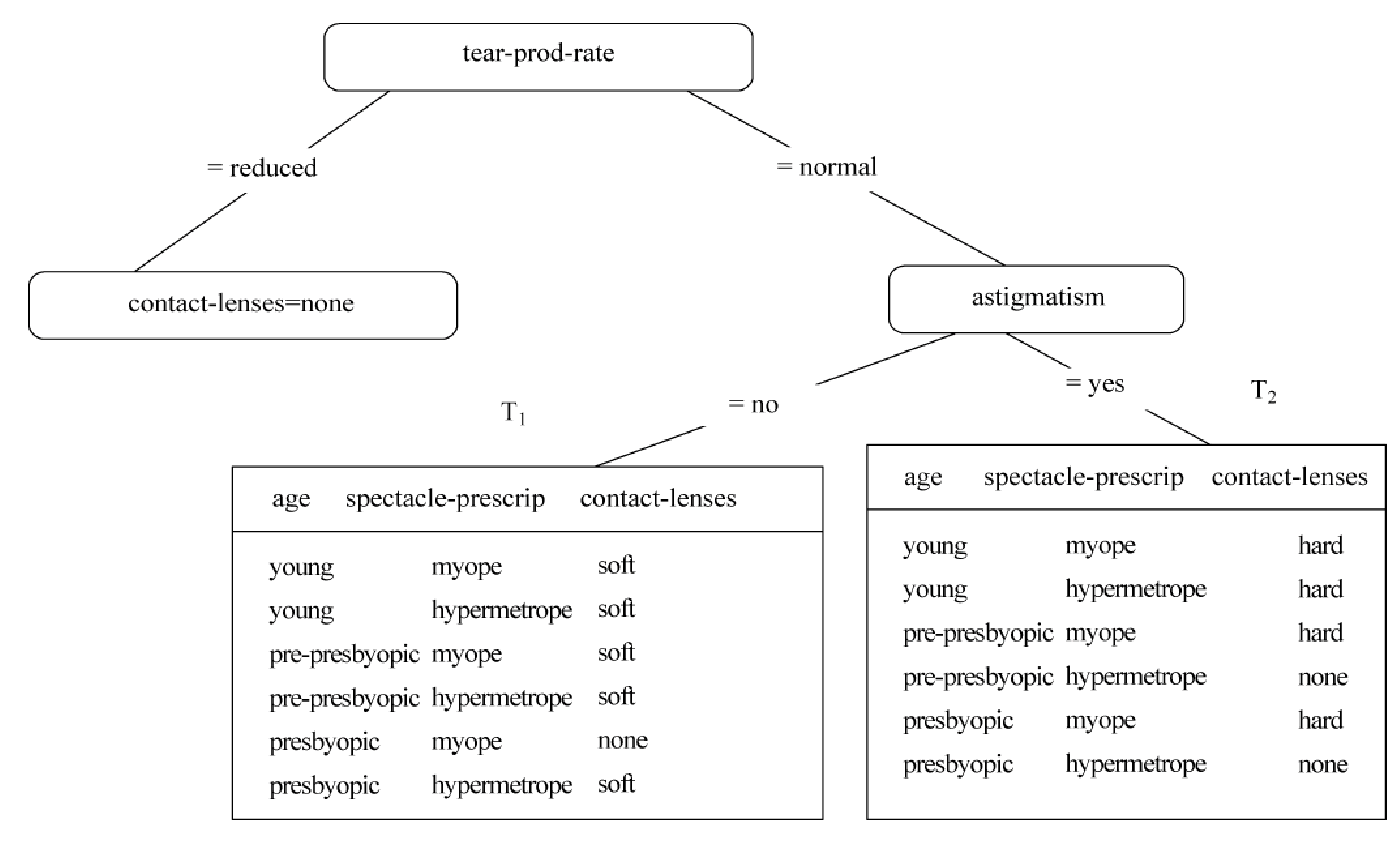


图 4-3 astigmatism 结点及其分支

对于 astigmatism=no 的左子树中的所有元组,由于仍然不能做出统一决策,因此需要进一步扩展。

首先计算出给定样本 contact-lenses 分类所需的期望信息:

$$I(s_1,s_2,s_3) = I(1,5,0) = -\frac{1}{6}\log_2\frac{1}{6} - \frac{5}{6}\log_2\frac{5}{6} - \frac{0}{6}\log_2\frac{0}{6}$$

$$= 0.431 + 0.219 + 0 = 0.65。$$

对于 astigmatism=no 的左子树中的所有元组(对应图 4-3 中的 T_1),计算其他两个属性的信息增益。

观察 age 的每个样本值 young、pre-presbyopic、presbyopic 的分布,具体情况如表 4-9 所示。

表 4-9 age 的样本值分布

contact-lenses / age	none	soft	hard
young	0	2	0
pre-presbyopic	0	2	0
presbyopic	1	1	0

对于 age=young，$s_{11}=0, s_{21}=2, s_{31}=0$，

$$I(s_{11}, s_{21}, s_{31}) = I(0,2,2) = -\frac{0}{2}\log_2\frac{0}{2} - \frac{2}{2}\log_2\frac{2}{2} - \frac{0}{2}\log_2\frac{0}{2} = 0+0+0 = 0;$$

对于 age=pre-presbyopic，$s_{12}=0, s_{22}=2, s_{32}=0$，

$$I(s_{12}, s_{22}, s_{32}) = I(0,2,0) = -\frac{0}{2}\log_2\frac{0}{2} - \frac{2}{2}\log_2\frac{2}{2} - \frac{0}{2}\log_2\frac{0}{2} = 0+0+0 = 0;$$

对于 age=presbyopic，$s_{13}=1, s_{23}=1, s_{33}=0$，

$$I(s_{13}, s_{23}, s_{33}) = I(1,1,0) = -\frac{1}{2}\log_2\frac{1}{2} - \frac{1}{2}\log_2\frac{1}{2} - \frac{0}{2}\log_2\frac{0}{2} = 0.5+0.5+0 = 1。$$

因此，如果样本按 age 划分，对一个给定的样本分类对应的熵和信息增益为：

$$E(\text{age}) = \frac{2}{6}I(s_{11}, s_{21}, s_{31}) + \frac{2}{6}I(s_{12}, s_{22}, s_{32}) + \frac{2}{6}I(s_{13}, s_{23}, s_{33})$$

$$= \frac{1}{3}(0+0+1) = 0.33;$$

$$Gain(\text{age}) = I(s_1, s_2, s_3) - E(age) = 0.65 - 0.33 = 0.32。$$

观察 spectacle-prescrip 的每个样本值 myope、hypermetrope 的分布（对应图 4-3 中 T_1），具体情况如表 4-10 所示。

表 4-10　spectacle-prescrip 的样本值分布

spectacle-prescrip \ contact-lenses	none	soft	hard
myope	1	2	0
hypermetrope	0	3	0

对于 spectacle-prescrip=myope，$s_{11}=1, s_{21}=2, s_{31}=0$，

$$I(s_{11}, s_{21}, s_{31}) = I(1,2,0) = -\frac{1}{3}\log_2\frac{1}{3} - \frac{2}{3}\log_2\frac{2}{3} - \frac{0}{3}\log_2\frac{0}{3}$$

$$= 0.53 + 0.39 + 0 = 0.92;$$

对于 spectacle-prescrip=hypermetrope，$s_{12}=0, s_{22}=3, s_{32}=0$，

$$I(s_{12}, s_{22}, s_{32}) = I(0,3,0) = -\frac{0}{3}\log_2\frac{0}{3} - \frac{3}{3}\log_2\frac{3}{3} - \frac{0}{3}\log_2\frac{0}{3} = 0+0+0 = 0。$$

因此，如果样本按 spectacle-prescrip 划分，对一个给定的样本分类对应的熵和信息增益为：

$$E(\text{spectacle-prescrip}) = \frac{3}{6}I(s_{11}, s_{21}, s_{31}) + \frac{3}{6}I(s_{12}, s_{22}, s_{32}) = \frac{1}{2}(0.92+0) = 0.46;$$

$$Gain(\text{spectacle-prescrip}) = I(s_1, s_2, s_3) - E(\text{spectacle-prescrip}) = 0.65 - 0.46 = 0.19。$$

由于 age 在属性中具有最高的信息增益，所以它被选作测试属性。并以此创建一个结点，用 age 标记，并对于每个属性值，引出一个分支，数据集被划分成三个子集。图 4-4 给出了 age 结点及其分支。

针对图 4-4 中的 T_1，可得出 age=young，contact-lenses=soft。

针对图 4-4 中的 T_2，可得出 age=pre-presbyopic，contact-lenses=soft。

针对图 4-4 中的 T_3，可继续划分，得出 spectacle-prescrip=myope 的情况下 contact-

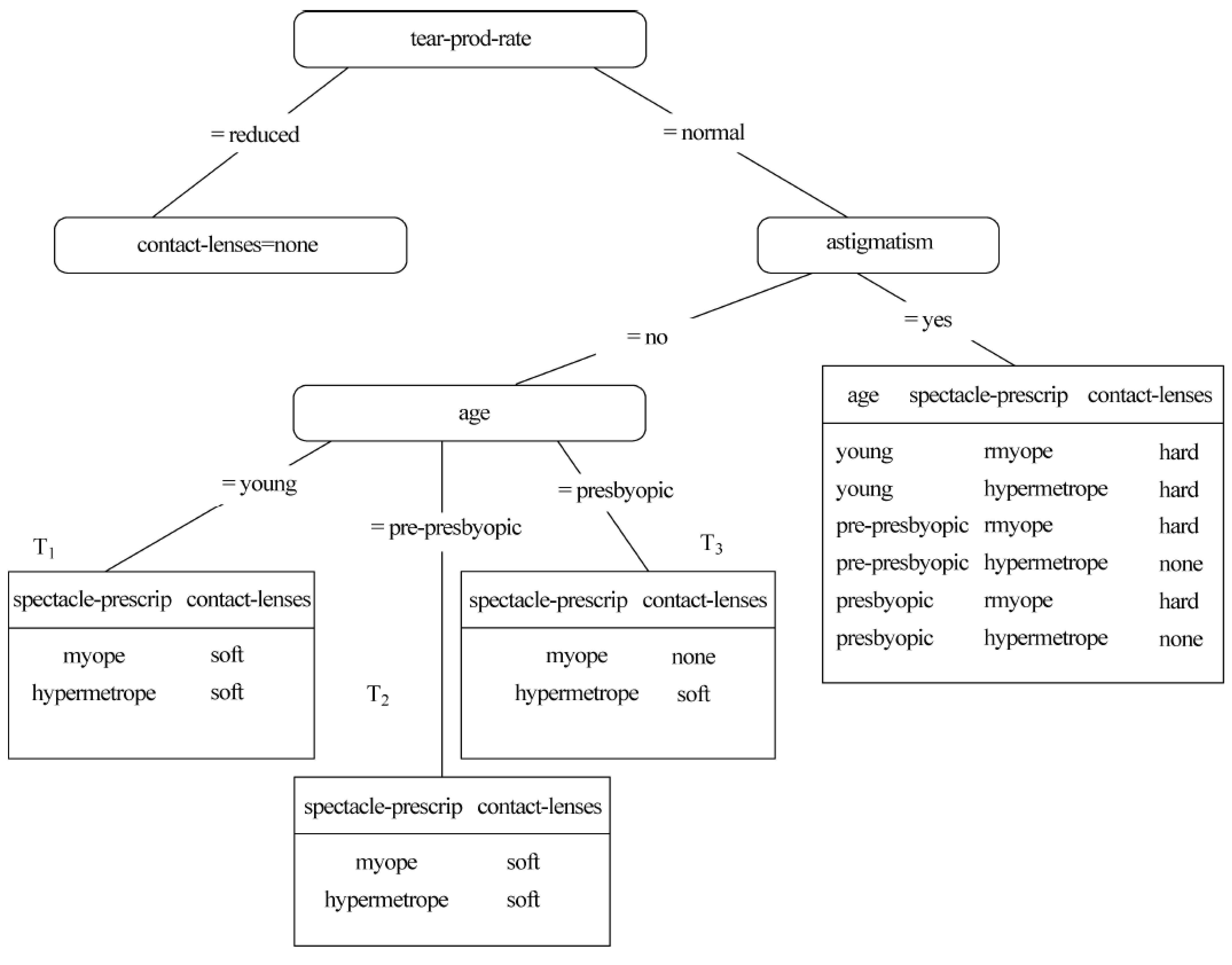


图 4-4　age 结点及其分支

lenses=none,spectacle-prescrip=hypermetrope 的情况下 contact-lenses=soft。

对于 astigmatism=yes 的右子树中的所有元组,由于仍然不能做出统一决策,因此需要进一步扩展。对于 astigmatism=yes 的右子树中的所有元组(对应图 4-3 中的 T_2),首先计算出给定样本 contact-lenses 分类所需的期望信息:

$$I(s_1,s_2,s_3)=I(2,0,4)=-\frac{2}{6}\log_2\frac{2}{6}-\frac{0}{6}\log_2\frac{0}{6}-\frac{4}{6}\log_2\frac{4}{6}$$
$$=0.53+0+0.39=0.92。$$

对于 astigmatism=yes 的右子树,计算其他两个属性的信息增益:

观察 age 的每个样本值 young、pre-presbyopic、presbyopic 的分布,具体情况如表 4-11 所示。

表 4-11　age 的样本值分布

age \ contact-lenses	none	soft	hard
young	0	0	2
pre-presbyopic	1	0	1
presbyopic	1	0	1

对于 age=young，$s_{11}=0$，$s_{21}=0$，$s_{31}=2$，

$$I(s_{11},s_{21},s_{31}) = I(0,2,2) = -\frac{0}{2}\log_2\frac{0}{2}-\frac{0}{2}\log_2\frac{0}{2}-\frac{2}{2}\log_2\frac{2}{2}=0+0+0=0;$$

对于 age=pre-presbyopic，$s_{12}=1$，$s_{22}=0$，$s_{32}=1$，

$$\begin{aligned}I(s_{12},s_{22},s_{32}) &= I(0,2,0) = -\frac{1}{2}\log_2\frac{1}{2}-\frac{0}{2}\log_2\frac{0}{2}-\frac{1}{2}\log_2\frac{1}{2}\\ &=0.5+0+0.5=1;\end{aligned}$$

对于 age=presbyopic，$s_{13}=1$，$s_{23}=0$，$s_{33}=1$，

$$\begin{aligned}I(s_{13},s_{23},s_{33}) &= I(1,0,1) = -\frac{1}{2}\log_2\frac{1}{2}-\frac{0}{2}\log_2\frac{0}{2}-\frac{1}{2}\log_2\frac{1}{2}\\ &=0.5+0+0.5=1。\end{aligned}$$

因此，如果样本按 age 划分，对一个给定的样本分类对应的熵和信息增益为：

$$E(\text{age}) = \frac{2}{6}I(s_{11},s_{21},s_{31})+\frac{2}{6}I(s_{12},s_{22},s_{32})+\frac{2}{6}I(s_{13},s_{23},s_{33})=\frac{1}{3}(0+1+1)=0.67;$$

$$Gain(age) = I(s_1,s_2,s_3)-E(age)=0.92-0.67=0.25。$$

观察 spectacle-prescrip 的每个样本值 myope、hypermetrope 的分布，具体情况如表 4-12 所示。

表 4-12 spectacle-prescrip 的样本值分布

spectacle-prescrip \ contact-lenses	none	soft	hard
myope	0	0	3
hypermetrope	2	0	1

对于 spectacle-prescrip=myope，$s_{11}=0$，$s_{21}=0$，$s_{31}=3$，

$$I(s_{11},s_{21},s_{31}) = I(0,0,3) = -\frac{0}{3}\log_2\frac{0}{3}-\frac{0}{3}\log_2\frac{0}{3}-\frac{3}{3}\log_2\frac{3}{3}=0+0+0=0;$$

对于 spectacle-prescrip=hypermetrope，$s_{12}=2$，$s_{22}=0$，$s_{32}=1$，

$$\begin{aligned}I(s_{12},s_{22},s_{32}) &= I(0,3,0) = -\frac{2}{3}\log_2\frac{2}{3}-\frac{0}{3}\log_2\frac{0}{3}-\frac{1}{3}\log_2\frac{1}{3}\\ &=0.53+0+0.39=0.92。\end{aligned}$$

因此，如果样本按 spectacle-prescrip 划分，对一个给定的样本分类对应的熵和信息增益为：

$$\begin{aligned}E(\text{spectacle-prescrip}) &= \frac{3}{6}I(s_{11},s_{21},s_{31})+\frac{3}{6}I(s_{12},s_{22},s_{32})\\ &=\frac{1}{2}(0.92+0)=0.46;\end{aligned}$$

$$\begin{aligned}Gain(\text{spectacle-prescrip}) &= I(s_1,s_2,s_3)-E(\text{spectacle-prescrip})\\ &=0.92-0.46=0.46。\end{aligned}$$

由于 spectacle-prescrip 在属性中具有最高的信息增益，所以它被选作测试属性。并以此创建一个结点，用 spectacle-prescrip 标记，并对于每个属性值，引出一个分支，数据集被划分成两个子集。图 4-5 给出了 spectacle-prescrip 结点及其分支。

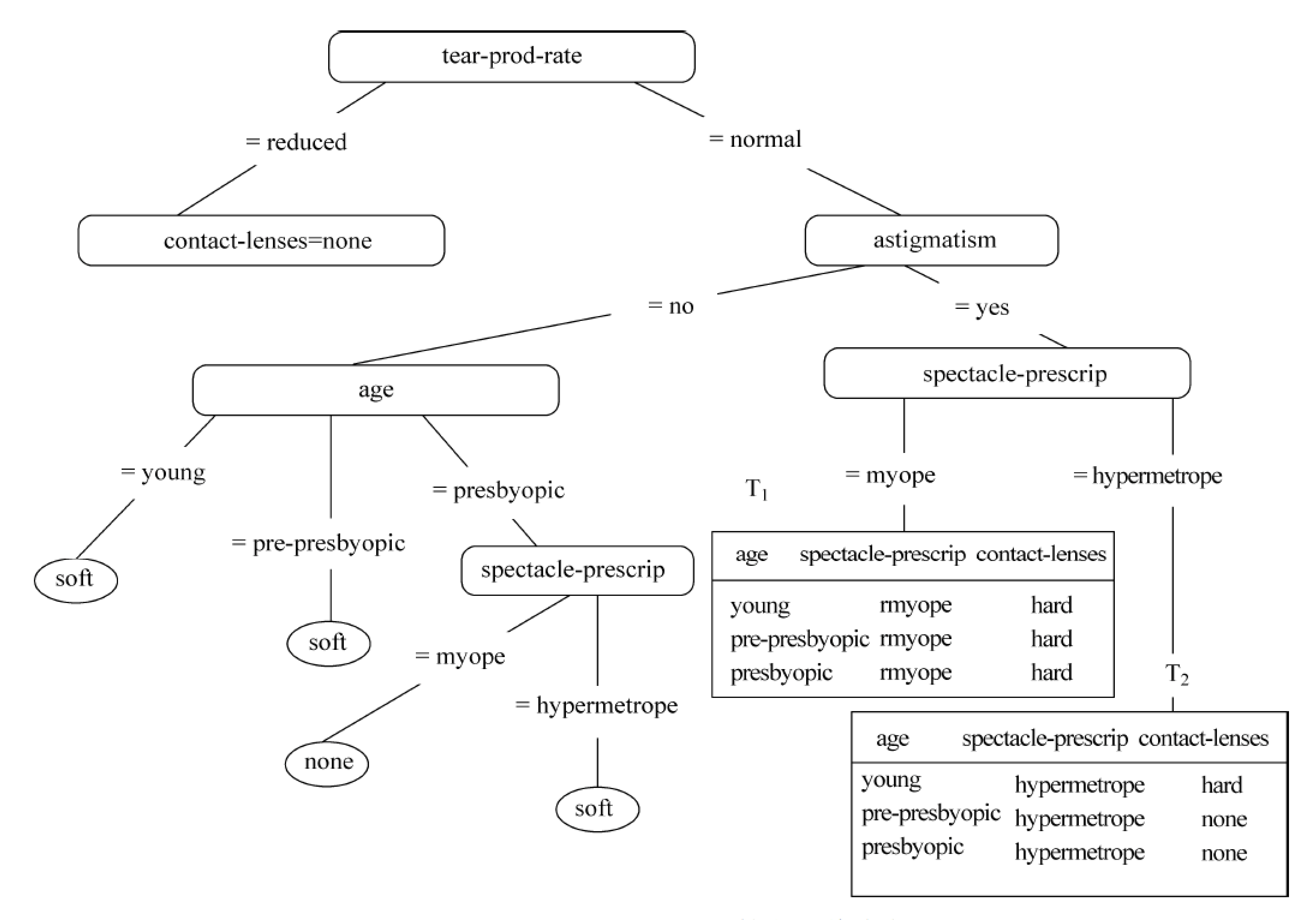


图 4-5 spectacle-prescrip 结点及其分支

针对图 4-5 中的 T_1，可得出 contact-lenses＝hard。

针对图 4-5 中的 T_2，可继续划分，得出 age＝young 的情况下 contact-lenses＝hard；age＝pre-presbyopic 的情况下 contact-lenses＝none；得出 age＝presbyopic 的情况下 contact-lenses＝none。

因此，最终的决策树如图 4-6 所示。

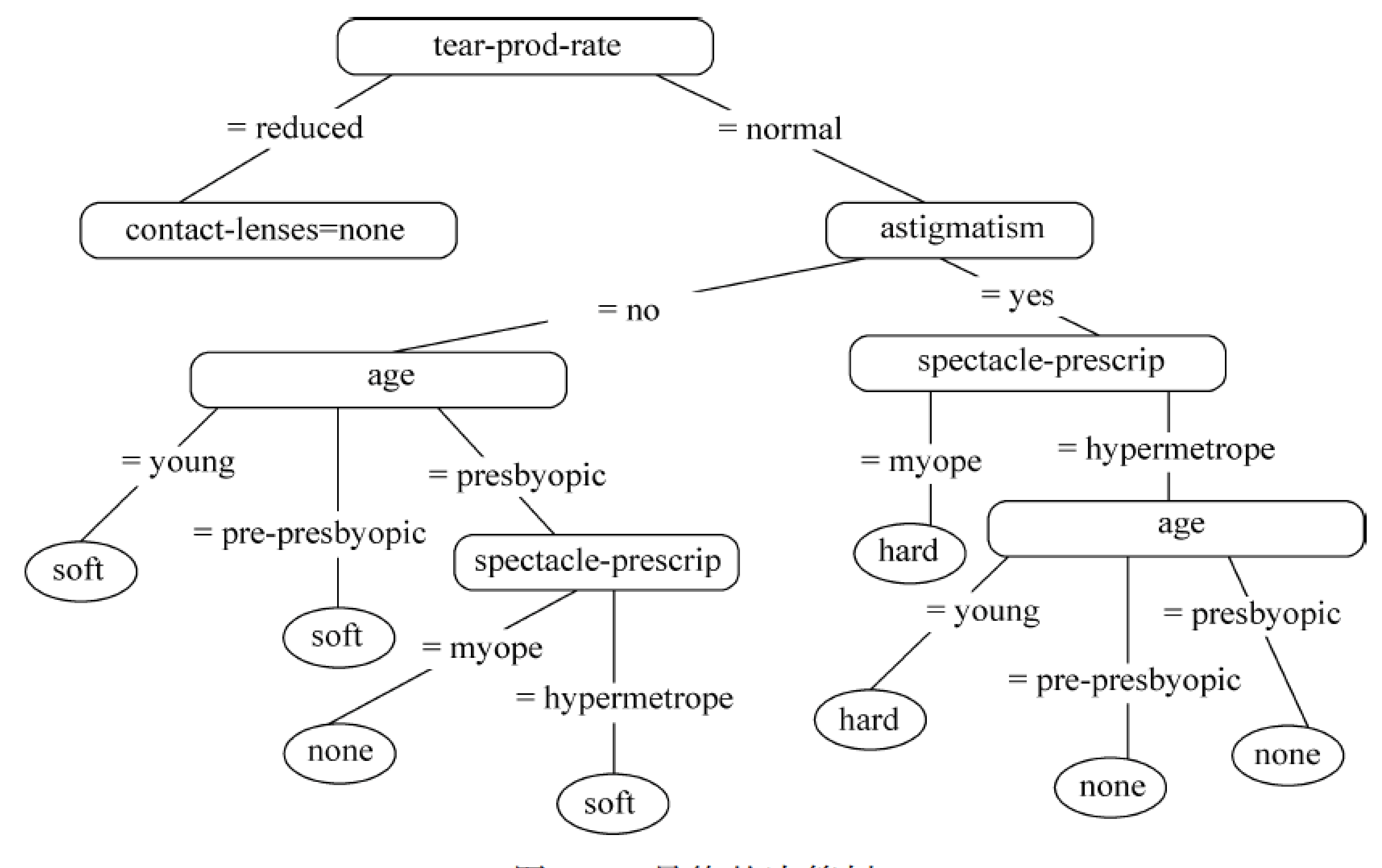


图 4-6 最终的决策树

10．用程序实现 ID3 算法，并测试上题的结果。

参考答案：略。

11．下面的例子被分为 3 类：{Short，Tall，Medium}，Height 属性被划分为(0，1.6)，(1.6，1.7)，(1.7，1.8)，(1.8，1.9)，(1.9，2.0)，(2.0，∞)，通过表 4-13，请用贝叶斯分类方法对例子 t=<Adam，M，1.95m>进行分类。

表 4-13　训练数据集

No.	Name	Gender	Height	Output
1	Kristina	F	1.6m	Short
2	Jim	M	2m	Tall
3	Maggie	F	1.9m	Medium
4	Martha	F	1.88m	Medium
5	Stephanie	F	1.7m	Short
6	Bob	M	1.85m	Medium
7	Kathy	F	1.6m	Short
8	Dave	M	1.7m	Short
9	Worth	M	2.2m	Tall
10	Steven	M	2.1m	Tall
11	Debbie	F	1.8m	Medium
12	Todd	M	1.95m	Medium
13	Kim	F	1.9m	Medium
14	Amy	F	1.8m	Medium
15	Wynette	F	1.75m	Medium

参考答案：标号类属性 Output 具有 3 个不同值{Short，Tall，Medium}。设 C1 对应于类 Output="Short"，C2 对应于类 Output="Medium"，C3 对应于类 Output="Tall"。希望分类的未知样本 t=<Adam，M，1.95m>，因此需要最大化 P(X|Ci)P(Ci)，i=1，2，3。

每个类的先验概率 P(Ci)可以根据训练样本计算：

- P(Output="Short")=4/15=0.267；
- P(Output="Medium")=8/15=0.533；
- P(Output="Tall")=3/15=0.200。

为计算 P(X|Ci)，i=1，2，3，计算下面的条件概率：

- P(Gender="M"|Output="Short")=1/4=0.25；
- P(Gender="M"|Output="Medium")=2/8=0.25；
- P(Gender="M"|Output="Tall")=3/3=1。
- P(Height=(1.9，2.0]|Output="Short")=0/4=0；
- P(Height=(1.9，2.0]|Output="Medium")=1/8=0.125；
- P(Height=(1.9，2.0]|Output="Tall")=1/3=0.33。

假设条件独立性，使用以上概率，得到：

- P(X|Output="Short")=0.25×0=0；
- P(X|Output="Medium")=0.25×0.125=0.0313；

- P(X|Output="Tall")=1×0.33=0.33。
- P(X|Output="Short")P(Output="Short")=0×0.267=0；
- P(X|Output="Medium")P(Output="Medium")=0.0313×0.533=0.0167；
- P(X|Output="Tall")P(Output="Tall")=0.33×0.2=0.066。

因此，对于样本 t=＜Adam,M,1.95m)，朴素贝叶斯分类预测 Output="Tall"。

12. 在应用贝叶斯方法解决实际问题的时候，可能会出现观察概率为 0 的情况，因此在贝叶斯分类中这项概率占有统治地位，如何解决上述问题？

参考答案：略。

13. EM 算法分为哪两个主要步骤？

参考答案：在 EM 算法的一般形式里，它重复以下两个步骤：E 步骤和 M 步骤，直至收敛。

(1) 估计(E)步骤：使用当前假设 h 和观察到的数据 X 来估计 Y 上的概率分布以计算 $Q(h'|h)$：$Q(h'|h)\leftarrow E[\ln P(Y|h')|h,X]$

(2) 最大化(M)步骤：将假设 h 替换为使 Q 函数最大化的假设 h'：$h\leftarrow \arg\max_{h'} Q(h'|h)$。

14. 简述 EM 算法每个步骤的主要作用。

参考答案：略。

15. 简述 AQ 算法中“种子”与“星”的概念。

参考答案：AQ 算法中“种子”是一个正例，“星”是覆盖种子而同时排除所有反例的概念描述或规则。

16. 假设有一个训练集，其包含三个属性：at1，at2，at3。现有正例负例样本分别如表 4-14 和表 4-15 所示，请用 AQ 算法对+类的规则进行获取。

表 4-14 正例样本

at1	at2	at3	class
y	n	r	+
x	m	r	+
y	n	s	+
x	n	r	+

表 4-15 负例样本

at1	at2	at3	class
x	m	s	−
y	m	t	−
y	n	t	−
z	n	t	−
z	n	r	−
x	n	s	−

17. 与 ID3 算法相比，CN2 算法有哪些特点？

参考答案：ID3 是典型的应用信息增益进行决策树分析的分类算法。CN2 算法结合了

ID3算法处理数据的效率和处理噪音数据的能力，以及AQ算法家族的灵活性。通过改进去除了对特定数据的依赖，且通过统计学类比，它可以达到与使用树剪枝方法的算法同样的效果。CN2使用一种基于噪音估计的启发式方法来终止它的搜索过程。使用这种方法可以不用对所有的训练样本进行正确的区分，但是规约出的规则在对新数据的处理上有很好的表现。

18. 假设有一个训练集，用CN2算法对上面第16题给出的数据集进行分析，找出相应的规则。

参考答案：略。

19. 简述FOIL算法的主要特点。

参考答案：FOIL用来对无约束的一阶Horn字句进行学习。FOIL算法由一个空子句开始查找，其不断地向当前的子句中追加文字直到没有负样例被子句所覆盖。之后，FOIL重新开始一个子句的查找，直到所有的正样例均被已经生成的子句所覆盖。

20. 简述FOIL算法与CN2算法的主要不同点。

参考答案：略。

21. 简述分类数据预处理的主要方法。

参考答案：数据预处理的主要方法如下：

(1) 数据清理：主要是消除或减少数据噪声和处理空缺值。

(2) 特征选择：从已知一组特征集中按照某一准则选择出有很好的区分特性的特征子集，或按照某一准则对特征的分类性能进行排序，用于分类器的优化设计。

(3) 数据变换：通过平滑、聚集、数据概化、规范化、特征构造等手段将数据转化为适合于挖掘的形式。

22. 简述分类中数据清理的常用方法。

参考答案：略。

23. 简述分类器的性能表示与评估的主要方法。

参考答案：分类器性能的表示方法类似信息检索系统的评价方法，可以采用OC曲线和ROC曲线、混淆矩阵等。

常用的评估分类方法有保持法和交叉验证两种主要方法。

(1) 保持法：把给定的数据随机地划分成训练集和测试集这两个独立的集合。通常，三分之一的数据分配到训练集，三分之二的数据分配到测试集。使用训练集得到分类器，其准确率用测试集评估。

(2) 交叉验证：把数据随机地分成不相交的、大小基本相等的n份。从这n份数据中抽取1份出来用作模型测试，其余$n-1$份数据合在一起建立模型，用先抽取出来的那1份数据对此模型做测试。这个过程对每一份数据都重复一次，即训练和测试都进行n次，得到n个不同的错误率，最后用所有数据建立一个模型，模型的错误率就是上述n个错误率的平均。

24. 如何评价分类器的性能？

参考答案：略。

第5章 聚类方法

1. 简单地描述下列英文缩写或短语的含义。

(1) Partitioning Method

(2) Hierarchical Method

(3) Density-based Method

(4) Grid-based Method

参考答案:(1) 划分法。它将数据划分为 k 个组,同时满足如下的要求:每个组至少包含一个对象;每个对象必须属于且只属于一个组。

(2) 层次法。它是对给定数据对象集合进行层次的分解。其基本思想是将模式样本按距离准则逐步聚类,直到满足分类要求为止。根据层次的分解如何形成,层次的方法又可以分为凝聚的和分裂的。

(3) 基于密度的方法。它将具有相同密度域的连通区域作为一簇。因此,它需要扫描整个数据集,将数据空间划分为不同的小方格,并使用小方格的并来近似表示簇。

(4) 基于网格的方法。这种方法首先将数据空间划分成为有限个单元(Cell)的网格结构,所有的处理都是以单个单元为对象的。这样处理的一个突出优点是处理速度快,通常与目标数据库中记录的个数无关,只与把数据空间分为多少个单元有关。

2. 简单地描述下列英文缩写或短语的含义。

(1) PAM

(2) STING

(3) DBSCAN

参考答案:略。

3. 简述聚类的基本概念。

参考答案:聚类就是把整个数据分成不同的组,并使组与组之间的差距尽可能大,组内数据的差异尽可能小。

聚类分析的输入可以用一组有序对(X,s)或(X,d)表示,这里 X 表示一组样本,s 和 d 分别是度量样本间相似度或相异度(距离)的标准。聚类系统的输出是一个分区,若$C=\{C_1,C_2,\cdots,C_k\}$,其中 $C_i(i=1,2\cdots,k)$是 X 的子集,有:

$$C_1 \bigcup C_2 \bigcup,\cdots,\bigcup C_k=X$$

$$C_1 \bigcap C_2=\varnothing,\quad i\neq j$$

C 中的成员 $C_1,C_2,\cdots,C_k$ 叫做类。

4. “物以类聚,人以群分”,请举例说明聚类的基本概念。

参考答案:略。

5. 聚类分析具有重要的作用,简述聚类分析在数据挖掘中的应用。

参考答案：聚类分析在数据挖掘中的应用主要有以下几个方面：

（1）聚类分析可以作为其他算法的预处理步骤。利用聚类进行数据预处理，可以获得数据的基本概况，在此基础上进行特征抽取或分类就可以提高精确度和挖掘效率。也可将聚类结果用于进一步关联分析，以获得进一步的有用信息。

（2）可以作为一个独立的工具来获得数据的分布情况。聚类分析是获得数据分布情况的有效方法。例如，在商业上，聚类分析可以帮助市场分析人员从客户基本库当中发现不同的客户群，并且用购买模式来刻画不同的客户群的特征。通过观察聚类得到的每个簇的特点，可以集中对特定的某些簇作进一步分析。这在诸如市场细分、目标顾客定位、业绩估评、生物种群划分等方面具有广阔的应用前景。

（3）聚类分析可以完成孤立点挖掘。许多数据挖掘算法试图使孤立点影响最小化，或者排除它们。然而孤立点本身可能是非常有用的。如在欺诈探测中，孤立点可能预示着欺诈行为的存在。

6. 举例说明聚类分析的用途。

参考答案：略。

7. 你认为一个好的聚类算法应该具备哪些特性？

参考答案：一个好的聚类算法应该具备如下特性：

- 可伸缩性；
- 处理不同类型属性的能力；
- 能够发现任意形状的聚类；
- 输入参数对领域知识的弱依赖性；
- 对于输入记录顺序不敏感；
- 挖掘算法应具有处理高维数据的能力；
- 处理噪声数据的能力；
- 基于约束的聚类；
- 挖掘出来的信息是可理解的和可用的。

8. 简述基于距离的聚类算法的主要特点。

参考答案：略。

9. 在对数据进行聚类的时候，会遇到二元特征样本，简述对二元特征样本进行距离度量的主要方法。

参考答案：假定 x 和 y 分别是 n 维特征，x_i 和 y_i 分别表示每维特征，且 x_i 和 y_i 的取值为二元类型数值$\{0,1\}$。则 x 和 y 的距离定义的常规方法是先求出如下几个参数，然后采用 SMC、Jaccard 系数或 Rao 系数。

（1）a 是样本 x 和 y 中满足 $x_i=y_i=1$ 的二元类型属性的数量。

（2）b 是样本 x 和 y 中满足 $x_i=1,y_i=0$ 的二元类型属性的数量。

（3）c 是样本 x 和 y 中满足 $x_i=0,y_i=1$ 的二元类型属性的数量。

（4）d 是样本 x 和 y 中满足 $x_i=y_i=0$ 的二元类型属性的数量。

(5) 简单匹配系数(Simple Match Coefficient,SMC)

$$S_{smc}(x,y)=\frac{a+b}{a+b+c+d}。$$

(6) Jaccard 系数

$$S_{jc}(x,y)=\frac{a}{a+b+c}。$$

(7) Rao 系数

$$S_{rc}(x,y)=\frac{a}{a+b+c+d}。$$

10. 哪种聚类算法对噪声数据不明显,可以发现不规则的类?

参考答案:略。

11. 给定两个对象,分别用元组(22,1,42,10),(20,0,36,8)表示。

(1) 计算两个对象之间的欧氏距离。

(2) 计算两个对象之间的绝对距离。

参考答案:(1) 根据两个对象之间的欧氏距离公式 $d(x,y)=\left[\sum_{i=1}^{n}|x_i-y_i|^2\right]^{1/2}$ 得出:

$$\begin{aligned}d(x,y)&=[\,|22-20|^2+|1-0|^2+|42-36|^2+|10-8|^2]^{1/2}\\&=(4+1+36+4)^{1/2}=45^{1/2}=6.708。\end{aligned}$$

(2) 根据两个对象之间的绝对距离公式 $d(x,y)=\sum_{i=1}^{n}|x_i-y_i|$ 得出:

$$d(x,y)=(22-20)+(1-0)+(42-36)+(10-8)=2+1+6+2=11。$$

12. 请说出在聚类分析中常用的距离度量方法。

参考答案:略。

13. 简述划分聚类方法的主要思想。

参考答案:给定一个有 n 个对象的数据集,划分聚类技术将构造数据 k 个划分,每一个划分就代表一个簇,$k\leqslant n$。也就是说,它将数据划分为 k 个簇,而且这 k 个划分满足下列条件:

- 每一个簇至少包含一个对象。
- 每一个对象属于且仅属于一个簇。

对于给定的 k,算法首先给出一个初始的划分方法,以后通过反复迭代的方法改变划分,使得每一次改进之后的划分方案都较前一次更好。所谓好的标准就是:同一簇中的对象越近越好,而不同簇中的对象越远越好。目标是最小化所有对象与其参照点之间的相异度之和。

14. 请说出划分聚类与层次聚类的主要特点。

参考答案:略。

15. 请用 k-平均算法把表 5-1 中的 8 个点聚为 3 个簇,假设第一次迭代选择序号 1、序号 4 和序号 7 当作初始点,请给出第一次执行后的三个聚类中心以及最后的三个簇。

表 5-1　样本数据 1

序号	属性 1	属性 2	序号	属性 1	属性 2
1	2	10	5	7	5
2	2	5	6	6	4
3	8	4	7	1	2
4	5	8	8	4	9

参考答案：对所给定的数据进行 k-平均算法(设 $n=8,k=3$)，以下为算法的执行步骤。

第一次迭代，假定随机选择的三个对象，如序号 1、序号 4 和序号 7 当作初始点，分别找到离三点最近的对象，并产生三个簇{1}、{3,4,5,6,8 }和{2,7}。

对于产生的簇分别计算平均值，得到平均值点。

- 对于{1}，平均值点为(2,10)；
- 对于{3,4,5,6,8 }，平均值点为(6,6)；
- 对于{2,7}，平均值点为(1.5,3.5)。

第二次迭代，通过平均值调整对象的所在的簇，重新聚类，即将所有点按离平均值点(2,10)、(6,6)、(1.5,3.5)最近的原则重新分配。得到三个新的簇：{1,8}、{3,4,5,6}和{2,7}。重新计算簇平均值点，得到新的平均值点为(3,9.5)、(6.5,5.25)和(1.5,3.5)。

第三次迭代，将所有点按离平均值点(3,9.5)、(6.5,5.25)和(1.5,3.5)最近的原则重新分配。得到三个新的簇：{1,4,8}、{3,5,6}和{2,7}。重新计算簇平均值点，得到新的平均值点为(3.67,9)、(7,4.33)和(1.5,3.5)。

第四次迭代，将所有点按离平均值点(3.67,9)、(7,4.33)和(1.5,3.5)最近的原则重新分配。调整对象，簇仍然为{1,4,8}、{3,5,6}和{2,7}，发现没有出现重新分配，而且准则函数收敛，程序结束。表 5-2 给出了整个过程中平均值计算和簇生成的过程和结果。

因此，第一次执行后的三个聚类中心为(2,10)、(6,6)、(1.5,3.5)，最后的三个簇为{1,4,8}、{3,5,6}，{2,7}。

表 5-2　平均值计算和簇生成的过程和结果

迭代次数	平均值(簇 1)	平均值(簇 2)	平均值(簇 3)	产生的新簇	新平均值(簇 1)	新平均值(簇 2)	新平均值(簇 3)
1	(2,10)	(5,8)	(1,2)	{1}，{3,4,5,6,8 }，{2,7}	(2,10)	(6,6)	(1.5,3.5)
2	(2,10)	(6,6)	(1.5,3.5)	{1,8}，{3,4,5,6}，{2,7}	(3,9.5)	(6.5,5.25)	(1.5,3.5)
3	(3,9.5)	(6.5,5.25)	(1.5,3.5)	{1,4,8}，{3,5,6}，{2,7}	(3.67,9)	(7,4.33)	(1.5,3.5)
4	(3.67,9)	(7,4.33)	(1.5,3.5)	{1,4,8}，{3,5,6}，{2,7}	(3.67,9)	(7,4.33)	(1.5,3.5)

16. 举例说明 k-平均算法的主要思想。

参考答案:略。

17. 请说出 k-平均算法的优点和缺点。

参考答案:k-平均算法的优点如下:

- k-平均算法简单、快速。
- 对处理大数据集,该算法是相对可伸缩的和高效率的,因为它的复杂度是 $O(n*k*t)$,其中,n 是所有对象的数目,k 是簇的数目,t 是迭代的次数。通常地,$k \ll n$,且 $t \ll n$。这个算法经常以局部最优结束。
- 算法尝试找出使平方误差函数值最小的 k 个划分。当结果簇是密集的,而簇与簇之间区别明显时,它的效果较好。

k-平均算法的缺点如下:

- k-平均方法只有在簇的平均值被定义的情况下才能使用。这可能不适用于某些应用,例如涉及有分类属性的数据。
- 要求用户必须事先给出 k(要生成的簇的数目),而且对初值敏感,对于不同的初始值,可能会导致不同的聚类结果。k-平均方法不适合于发现非凸面形状的簇,或者大小差别很大的簇。而且,它对于“噪声”和孤立点数据是敏感的,少量的该类数据能够对平均值产生极大影响。

18. 试比较 k-平均算法与 k-中心点算法的特点。

参考答案:略。

19. 简述 k-中心点算法的主要思路。

参考答案:k-中心点算法选用簇中位置最中心的对象作为代表对象,试图对 n 个对象给出 k 个划分。代表对象也被称为是中心点,其他对象则被称为非代表对象。最初随机选择 k 个对象作为中心点,该算法反复地用非代表对象来代替代表对象,试图找出更好的中心点,以改进聚类的质量。在每次迭代中,所有可能的对象对被分析,每个对中的一个对象是中心点,而另一个是非代表对象。对可能的各种组合,估算聚类结果的质量。一个对象 O_i 被可以产生最大平方-误差值减少的对象代替。在一次迭代中产生的最佳对象集合成为下次迭代的中心点。

20. 简述 PAM 算法的主要步骤。

参考答案:略。

21. 简述凝聚的层次聚类方法的主要思路。

参考答案:凝聚的层次聚类是一种自底向上的策略。首先将每个对象作为单独的一个簇,然后相继的合并相近的对象或组,将较小的数据对象子集合依据相似程度进行合并,这些小的数据对象子集合逐渐合并成较大的数据对象子集合,直到所有的类合并为一个,或者达到一个终止条件,从而构成一个簇的层次。

22. 在表 5-3 中给定的样本上运行 AGNES 算法,假定算法的终止条件为 3 个簇,初始簇{1},{2},{3},{4},{5},{6},{7},{8}。

表 5-3　样本数据 2

序号	属性 1	属性 2	序号	属性 1	属性 2
1	2	10	5	7	5
2	2	5	6	6	4
3	8	4	7	1	2
4	5	8	8	4	9

23. 在表 5-4 中给定的样本上运行 DIANA 算法，假定算法的终止条件为 3 个簇，初始簇{1,2,3,4,5,6,7,8}。

表 5-4　样本数据 3

序号	属性 1	属性 2	序号	属性 1	属性 2
1	2	10	5	7	5
2	2	5	6	6	4
3	8	4	7	1	2
4	5	8	8	4	9

参考答案：执行过程如下：

第 1 步，找到具有最大直径的簇，对簇中的每个点计算平均相异度（假定采用是欧氏距离）。

- 1 的平均距离：(5.000＋8.485＋3.606＋7.071＋7.211＋8.062＋2.236)/7＝5.953。
- 2 的平均距离：(5.000＋6.082＋4.243＋5.000＋4.123＋3.162＋4.472)/7＝4.583。
- 3 的平均距离：(8.485＋6.082＋5.000＋1.414＋2.000＋7.280＋6.403)/7＝5.238。
- 4 的平均距离：(3.606＋4.243＋5.000＋3.606＋4.123＋7.211＋1.414)/7＝4.172。
- 5 的平均距离：(7.071＋5.000＋1.414＋3.606＋1.414＋6.708＋5.000)/7＝4.316。
- 6 的平均距离：(7.211＋4.123＋2.000＋4.123＋1.414＋5.385＋5.385)/7＝4.234。
- 7 的平均距离：(8.062＋3.162＋7.280＋7.211＋6.708＋5.385＋7.616)/7＝6.489。
- 8 的平均距离：(2.236＋4.472＋6.403＋1.414＋5.000＋5.385＋7.616)/7＝4.647。

挑出平均相异度最大的点 7 放到 splinter group 中，剩余点在 old party 中。

第 2 步，在 old party 里找出到最近的 splinter group 中的点的距离不大于到 old party 中最近的点的距离的点，将该点放入 splinter group 中，该点是 2。

第 3 步，没有新的 old party 中的点被分配给 splinter group，此时分裂的簇数为 2。

第 4 步，此时具有最大直径的簇为{1,3,4,5,6,8}（具体属性值见表 5-5），对簇中的每个点计算平均相异度。

表 5-5　具有最大直径的簇对应的属性值

序号	属性 1	属性 2	序号	属性 1	属性 2
1	2	10	5	7	5
3	8	4	6	6	4
4	5	8	8	4	9

- 1 的平均距离:(8.485+3.606+7.071+7.211+2.236)/5=5.722。
- 3 的平均距离:(8.485+5.000+1.414+2.000+6.403)/5=4.660。
- 4 的平均距离:(3.606+5.000+3.606+4.123+1.414)/5=3.549。
- 5 的平均距离:(7.071+1.414+3.606+1.414+5.000)/5=3.701。
- 6 的平均距离:(7.211+2.000+4.123+1.414+5.385)/5=4.027。
- 8 的平均距离:(2.236+6.403+1.414+5.000+5.385)/5=4.088。

挑出平均相异度最大的点 1 放到 splinter group 中,剩余点在 old party 中。

第 5 步,没有新的 old party 的点被分配给 splinter group,此时分裂的簇数为 3,达到终止条件。表 5-6 给出了整个过程中平均值计算和簇生成的过程和结果。

表 5-6　平均值计算和簇生成的过程

步骤	具有最大直径的簇	splinter group	old party
1	{1,2,3,4,5,6,7,8}	{7}	{1,2,3,4,5,6,8}
2	{1,2,3,4,5,6,7,8}	{2,7}	{1,3,4,5,6,8}
3	{1,3,4,5,6,8}	{1}	{3,4,5,6,8}
4	{1,3,4,5,6,8}	{1}	{3,4,5,6,8}

24. 请分析 DIANA 和 AGNES 算法的特点。

参考答案:略。

25. 简述密度聚类方法的主要思路。

参考答案:密度聚类方法的指导思想是,只要一个区域中的点的密度大于某个阈值,就把它加到与之相近的聚类中去。

26. 请举例说明 DBSCAN 算法的主要思想。

参考答案:略。

27. 简述 STING 算法的主要特点。

参考答案:STING 算法是一种基于网格的多分辨率聚类技术,它将空间区域划分为矩形单元。由于存储在每个单元中的统计信息提供了单元中的数据不依赖于查询的汇总信息,因而计算是独立于查询的。

STING 算法的质量取决于网格结构最低层的粒度。如果粒度比较细,处理的代价会显著增加;但如果粒度较粗,则聚类质量会受到影响。

STING 算法的主要优点是效率高,通过对数据集的一次扫描来计算单元的统计信息,因此产生聚类的时间复杂度是 $O(n)$。在建立层次结构以后,查询的时间复杂度是 $O(g)$,g 远小于 n。此外,STING 算法采用网格结构,有利于并行处理和增量更新。

第 6 章　时间序列和序列模式挖掘

1. 简单地描述下列英文缩写或短语的含义。

(1) Sequential Mining

(2) Time Series

(3) Offset Translation

(4) Subsequence Ordering

参考答案：

(1) 序列挖掘。它是指从数据库中发现相对时间或者其他顺序出现的高频率子序列。

(2) 时间序列。从统计意义上来讲，所谓时间序列就是将某一指标在不同时间上的不同取值，按照时间的先后顺序排列而成的数列。

(3) 偏移变换。将数据使用偏移技术进行转换，以利于处理。

(4) 子序列排序。它的主要任务是从没有重叠的子序列匹配中找出匹配最长的那些序列。

2. 解释下列概念。

(1) 时间序列

(2) 偏移变换

(3) 多元时间序列

(4) 子序列匹配

参考答案：略。

3. 简述时间序列挖掘的概念。

参考答案：时间序列挖掘就是要从大量的时间序列数据中提取人们事先不知道的，但又是潜在有用的、与时间属性相关的信息和知识，并短期、中期和长期预测，指导人们的社会、经济、军事和生活等行为。

4. 举例说明时间序列挖掘的意义。

参考答案：略。

5. 简述时间序列预测的常用方法。

参考答案：时间序列预测的常用方法有：

(1) 确定性时间序列预测方法。设法消除随机型波动、分解季节性变化、拟合确定型趋势，因而形成对发展水平分析、趋势变动分析和长期趋势加周期波动分析等一系列确定性时间序列预测方法。

(2) 随机时间序列预测方法。通过建立随机模型，对随机时间序列进行分析，可以预测未来值。若时间序列是平稳的，可以用自回归模型、移动回归模型或自回归移动平均模型进行分析预测。

(3) 神经网络方法。通过对某段历史数据的训练，通过数学统计模型估计神经网络的各层权重参数初值，从而建立神经网络预测模型，用于时间序列的预测。

6. 简述常见的确定性时间序列预测模型。

参考答案:略。

7. ARMA 模型是时序方法中最基本的、实际应用最广的时序模型,请简述该模型的主要思想。

参考答案:由于 ARMA 模型是一个信息的凝聚器,可将系统的特性与系统状态的所有信息凝聚在其中,因而它也可以用于时间序列的匹配。AR 模型描述的是系统对过去自身状态的记忆,MA 模型描述的是系统对过去时刻进入系统的噪声的记忆,而 ARMA 模型则是系统对过去自身状态以及各时刻进入的噪声的记忆。

8. 请简述 AR 模型参数矩阵估计的方法,以及判别函数的构造方法。

参考答案:略。

9. 在时间序列分析方面,离散傅里叶变换具有独特的优点。请简述采用该方法进行完全匹配的主要思想。

参考答案:所谓完全匹配必须保证被查找的序列与给出的序列有相同的长度。首先进行特征提取,即对给定的时间序列进行离散傅里叶变换。其次进行首次筛选,用欧氏距离衡量两个序列是否相似的一般方法。如果两个序列的欧氏距离小于 ε 的话,则认为这两个序列相似;从提出特征后的频域空间中找出满足某一式子的序列,这样就滤掉一大批与给定序列的距离大于 ε 的序列。

10. 在时间序列分析方面,离散傅里叶变换具有独特的优点。请简述采用该方法进行完全匹配的主要思想。

参考答案:略。

11. 与基于距离的比较方法和基于傅里叶变换时间序列查找方法相比,基于规范变换的查找方法具有哪些优点?

参考答案:这种方法经过原子匹配与窗口缝合就找出了相似的子序列,通过对子序列排序完成了序列的相似查找,因此该方法不仅适用于完全匹配,而且适用于子序列匹配。另外,这种方法中过滤掉了一些 Gap,而且对序列作幅度缩放和偏移变换,所以该方法具有良好的鲁棒性,在算法的具体执行中用户可以设定 $\omega,\gamma,\varepsilon$,增加了算法的适用性。

12. 请比较各种时间序列分析方法的特点。

参考答案:略。

13. 给定序列数据库 D_T,请说明 D_T 上的频繁 k-序列的具体含义。

参考答案:支持度大于最小值程度的 k-序列,称为 D_T 上的频繁 k-序列。

14. 请举例说明序列的包含关系。

参考答案:略。

15. 简述序列模式挖掘的一般步骤。

参考答案:序列模式挖掘包括以下步骤:

(1) 排序阶段。对数据库进行排序,排序的结果是将原始的数据库转换成序列数据库。

(2) 大项集阶段。这个阶段要找出所有频繁的项集(即大项集)组成的集合 L。实际

上，也同步得到所有大 1-序列组成的集合，即{<l>|$l \in L$}。

(3) 转换阶段。在寻找序列模式的时候，要不断地检测一个给定的大序列集合是否包含于一个客户序列中。在转换完成的客户序列中，每条交易被其所包含的所有大项集所取代。如果一条交易不包含任何大项集，在转换完成的序列中它将不被保留。但是，在计算客户总数的时候，它仍将被计算在内。

(4) 序列阶段。利用转换后的数据库寻找频繁的序列，即大序列。

(5) 选最大阶段。在大序列集中找出最长序列。

16. 简述序列模式挖掘的各个步骤的主要任务。

参考答案：略。

17. 请简述 AprioriAll 算法的主要思想。

参考答案：AprioriAll 算法源于频繁集算法 Apriori，它把 Apriori 的基本思想扩展到序列挖掘中，也是一个多遍扫描数据库的算法。在每一遍扫描中都利用前一遍的大序列来产生候选序列，然后在完成对整个数据库的遍历后测试它们的支持度。在第一遍扫描中，利用大项目集阶段的输出来初始化大 1-序列的集合。在每次遍历中，从一个由大序列组成的种子集开始，利用这个种子集，可以产生新的潜在的大序列。在第一次遍历前，所有在大项集阶段得到的大 1-序列组成了种子集。

18. 请用 AprioriAll 算法在如表 6-1 所示的数据库例子中找出大序列，假定最小支持度为 40%。

表 6-1　序列数据库示例

3-Sequence	Support
<4,5,7>	2
<4,5,6>	2
<4,6,7>	3
<5,6,7>	2
<4,6,8>	2

参考答案：略。

19. AprioriSome 算法的执行过程可以分为两个步骤，请简述每个步骤的主要任务。

参考答案：AprioriSome 算法可以看作是 AprioriAll 算法的改进，具体过程分为两个阶段：

(1) 前推阶段。此阶段用于找出指定长度的所有大序列。在前推阶段中，只对特定长度的序列进行计数。

(2) 回溯阶段。此阶段用于查找其他长度的所有大序列。在这个阶段，对那些在前推阶段忽略的长度的序列进行计算。因为需要的是最大序列，所以可以在前推阶段就删除所有包含在其他大序列中的序列，那些序列不属于需要找的答案集。同时也删除在前推阶段找到的那些非最长的大序列。

20. 请用 AprioriSome 算法对 18 题给出的数据库例子中找出大序列，假定最小支持度为 40%。

参考答案：略。

21. 请简述GSP算法的主要思想。

参考答案：GSP算法类似于Apriori算法，大体分为候选集产生、候选集计数以及扩展分类三个阶段。与AprioriAll算法相比，GSP算法统计较少的候选集，并且在数据转化过程中不需要事先计算频繁集。

GSP算法主要包括三个步骤：

(1) 扫描序列数据库，得到长度为1的序列模式L_1，作为初始的种子集；

(2) 根据长度为i的种子集L_i通过连接操作和剪切操作生成长度为$i+1$的候选序列模式C_{i+1}，然后扫描序列数据库，计算每个候选序列模式的支持数，产生长度为$i+1$的序列模式L_{i+1}，并将L_{i+1}作为新的种子集；

(3) 重复第二步，直到没有新的序列模式或新的候选序列模式产生为止。

22. 与AprioriSome和AprioriAll相比，GSP算法具有哪些优点？

参考答案：略。

第7章　Web挖掘技术

1. 简单地描述下列英文缩写或短语的含义。

(1) Web Content Mining

(2) Web Usage Mining

(3) Web Structure Mining

(4) Crawler

(5) Look up Page

参考答案：(1) Web内容挖掘。是对站点的Web页面的文本以及多媒体等内容进行的分析和挖掘。

(2) Web访问信息挖掘。是对用户访问Web时在服务器方留下的访问记录等进行挖掘，发现用户的潜在访问模式。

(3) Web结构挖掘。是对Web页面之间的链接结构进行挖掘。

(4) 爬虫。一个搜索引擎用的网络爬行者，能够从一个链接到另外一个链接，遍历网络，且识别和阅读网页的程序。

(5) 查找页。帮助用户查找站点内的特定内容。

2. 解释下列概念。

(1) 爬虫

(2) 导航页

(3) 数据入口页

(4) 用户会话

(5) 权威页面

(6) 中心页面

参考答案：略。

3. 简述Web数据挖掘的意义。

参考答案：Web挖掘的实质就是从Web页面及其链接和用户对页面的访问中挖掘出用户感兴趣的知识。通过Web数据挖掘，可以从数以亿计存储大量多种多样信息的Web页面及其链接和用户对页面的访问中挖掘出需要的有用知识。

4. 举例说明Web数据挖掘的意义。

参考答案：略。

5. 根据所挖掘的信息来源，Web数据挖掘可以分为哪几类？

参考答案：Web挖掘依靠它所挖掘的站点信息来源可以分为Web内容挖掘(Web Content Mining)、Web访问信息挖掘(Web Usage Mining)和Web结构挖掘(Web Structure Mining)三种主要类型。

6. 简述Web数据挖掘的分类，并对每类的主要任务进行描述。

参考答案：略。

7. 从基于关键词查询的搜索引擎存在的主要问题角度说明 Web 挖掘的必要性。

参考答案:基于关键词查询的搜索引擎至少有两个问题不可回避:

(1) 由于精确度低,使得搜索的结果的可用性大打折扣。

(2) 搜索结果是凌乱的、无组织的,因而无法反复使用。

Web 挖掘则是从 Web 页面及其链接和用户对页面的访问中挖掘出用户感兴趣的知识。因此,Web 挖掘有望解决目前搜索引擎存在的问题。

8. 如何理解 Web 挖掘是一个交叉研究的领域。

参考答案:略。

9. Web 挖掘的数据来源有哪些?

参考答案:Web 挖掘面向的是网站数据,这些数据包括网页文本信息、网页链接信息、网站的访问记录以及其他可收集的信息。但是,不同的挖掘目的、不同的挖掘算法总是依靠不同的一种或几种数据源。例如 Server 日志、Error 日志、Çookie 日志、在线市场数据、Web 页面、Web 页面超链接以及包括用户注册信息等数据源。

10. 举例说明 Web 挖掘可以对服务器日志数据进行挖掘。

参考答案:略。

11. Web 内容挖掘的目的是什么?

参考答案:Web 内容挖掘的目的之一是基于页面内容相似度进行用户分类或聚类的,个性化的建立是通过用户过去的检索内容分析而建立起来的。Web 内容挖掘目前主要可以用于权威页面的发现,以及分析相关的页面链接结构,并且通过分析这类信息来获取到更多需要的信息。例如,现在许多 Web 搜索引擎就利用 Web 内容挖掘中的 Web 超链分析算法来提高搜索的效率和准确性。

12. 为什么说 Web 内容挖掘的基本技术是文本挖掘?

参考答案:略。

13. Web 页面内容预处理的目的是什么?

参考答案:Web 页面内容预处理的目的是把包括文本(Text)、图片(Image)、Script 和其他一些多媒体文件所包含的信息转换成可以实施 Web 挖掘算法的规格化形式。

14. 举例说明 Web 内容挖掘在个性化方面的应用。

参考答案:略。

15. 简述 Web 访问信息挖掘的特点。

参考答案:Web 访问信息挖掘的特点:从挖掘对象的进一步领域化、对挖掘方法的要求以及挖掘目的三个角度说明 Web 访问信息挖掘的特殊性。

16. 与传统的基于事务的数据挖掘方法相比,Web 访问信息挖掘对象有哪些独特的特点?

参考答案:略。

17. Web 访问信息挖掘的意义是什么?

参考答案:Web 访问信息挖掘的意义可以概括为如下几点。

(1) 改进 Web 站点的效率。通过对用户访问信息的挖掘,得到大多数用户的访问习惯、爱好和其他有用信息,利用这些信息可以指导网站提供商改进站点结构和布局,吸引更多用户。

(2) 实现个性化推荐。随着互联网的普及和电子商务的发展,电子商务系统在为用户提供越来越多选择的同时,其结构也变得更加复杂,用户经常会迷失在大量的商品信息空间中,无法顺利找到自己需要的商品。在日趋激烈的竞争环境下,个性化服务是包括电子商务在内的网站提供商争取更多用户、防止用户流失以及实现市场目标的重要手段。

(3) 商业智能的发现。从过去的访问信息特性的挖掘,发现新的商业智能,用于指导改进服务和扩展新的赢利点。通过结合日志数据和市场数据可以和 CRM 管理结合,在诸如顾客吸引(Customer Attraction)、顾客保留(Customer Retention)、跨区销售(Cross Sales)、顾客离开(Customer Departure)等市场活动中找到相应的最佳对策。

(4) 发现导航模式。用户的导航模式是指群体用户对 Web 站点内的页面的浏览顺序模式。在电子商务环境下发现商业智能的关键是发现用户的导航模式。这种导航模式也是个性化推销的基础。

(5) 抽取访问信息特性。通过对客户端,服务器端,代理服务器端等不同用户访问信息的挖掘可以得到关于用户交互情况和导航情况的详细的信息。在此基础上可以提出模型,用于预测在一个给定站点上一个用户所访问的页面的概率分布。访问信息的特性可以被用于在 Web 服务器上开展伸缩性和负载均衡的研究等方面。

18. 举例 Web 访问信息挖掘的好处。

参考答案:略。

19. Web 访问信息挖掘的作用。

参考答案:Web 访问信息挖掘的好处主要有:

(1) 利用 Web 访问信息挖掘可以实现用户建模;

(2) 利用 Web 访问信息挖掘发现导航模式,从而改进 Web 站点的结构设计,实行个性化推销;

(3) 利用 Web 访问信息挖掘改进访问效率,改进服务器的性能;

(4) 利用 Web 访问信息挖掘还可以进行个性化服务;

(5) 利用 Web 访问信息挖掘进行商业智能发现;

(6) 利用 Web 访问信息挖掘进行用户移动模式发现。

20. Web 访问信息挖掘的基础和最烦琐的工作是数据的预处理,请说出常用的 Web 访问信息挖掘的预处理方法。

参考答案:略。

21. Web 访问信息挖掘中的常用技术有哪些?

参考答案:Web 访问信息挖掘中的常用技术有如下几种。

(1) 路径分析。路径分析最常见的应用是用于判定在一个 Web 站点中最频繁访问的路径,这样的知识对于一个电子商务网站或者信息安全评估是非常重要的。

(2) 关联规则发现。使用关联规则发现方法可以从 Web 访问事务集中,找到一般性的关联知识。

（3）序列模式发现。在时间戳有序的事务集中，序列模式的发现就是指找到那些如"一些项跟随另一个项"这样的内部事务模式。

（4）分类。发现分类规则可以给出识别一个特殊群体的公共属性的描述。

（5）聚类。可以从 Web Usage 数据中聚集出具有相似特性的那些客户。

22．举例说明 Web 访问信息挖掘中可采用的挖掘方法。

参考答案：略。

23．请解释用户建模，并说出常见的用户建模方法。

参考答案：用户建模（Modelling Users）是指根据访问者对一个 Web 站点上 Web 页面的访问情况，模型化用户的自身特性。在识别出用户的特性后就可以开展针对性的服务。

常见方法有：

（1）推断匿名访问者的人口统计特性；

（2）在不打扰用户的情况下，得到用户概貌文件；

（3）根据用户的访问模式来聚类用户。

24．Web 访问信息挖掘可以实现用户建模，请比较各种用户建模方法。

参考答案：略。

25．简述利用 Web 访问信息挖掘发现导航模式的意义。

参考答案：发现导航模式（Discovering Navigation Patterns）是 Web 访问信息挖掘的一个重要的研究领域。用户的导航模式是指群体用户对 Web 站点内的页面的浏览顺序模式。用户导航模式的主要应用在改进站点设计和个性化推销等方面。得到的导航模式可以指导网站设计人员改进站点的设计结构，吸引用户的访问，在电子商务环境下发现市场智能的关键是发现用户的导航模式，这种导航模式可以被用于个性化的推销。

26．发现导航模式是 Web 访问信息挖掘的一个重要的研究领域，请简单介绍一些比较著名的导航模式发现方法。

参考答案：略。

27．为什么 Web 访问信息挖掘能够改进访问效率？

参考答案：利用 Web 访问信息挖掘结果可以在许多方面改进 Web 站点的访问效率，Web 服务器推送技术，自适应网站，利用导航模式的结果改进 Web 服务器的性能这些都能改进访问效率，而这些技术的改进都可以通过访问信息挖掘。

28．请举例说明 Web 访问信息挖掘能够改进访问效率。

参考答案：略。

29．请简述在 Web 站点开展个性化服务的总体思路和步骤。

参考答案：在 Web 站点开展个性化（Personalization）服务的总的思路和步骤是：

（1）模型化页面和用户；

（2）分类页面和用户；

（3）在页面和对象之间进行匹配；

（4）判断当前访问的类别以进行推荐。

30. 举例说明 Web 访问信息挖掘在个性化服务方面的应用。

参考答案：略。

31. 请简述 Web 结构挖掘的主要任务和目的。

参考答案：在设计搜索引擎等服务时，对 Web 页面的链接结构进行挖掘以得出有用的知识是提高检索效率的重要手段。Web 页面的链接类似学术上的引用，因此一个重要的页面可能会有很多页面的链接指向它。也就是说，如果有很多链接指向一个页面，那么它一定是重要的页面。通过链接结构的挖掘，来发现这些重要页面。

32. 请给出一种 Web 站点遍历的思路。

参考答案：略。

第8章　空间挖掘

1. 简单地描述下列英文缩写或短语的含义。

(1) Spatial Mining

(2) Spatial Statistics

(3) Minimum Bounding Rectangle

(4) Geographic Information System

(5) Spatial Online Analytical Mining

参考答案:(1) 空间挖掘。通常被称作空间数据挖掘,或者空间数据库的知识发现,是数据挖掘在空间数据库或空间数据方面的应用。

(2) 空间统计学。是依靠有序的模型来描述无序事件,根据不确定性和有限的信息来分析、评价和预测空间数据。

(3) 最小包围矩形(MBR)。指能够包围某一个图形的面积最小的矩形。

(4) 地理信息系统(GIS)。是以地理空间数据库为基础,对空间数据进行采集、储存、管理、分析、模拟和显示,实时提供空间和动态的地理环境信息,并服务于辅助决策的空间信息系统。

(5) 空间在线分析挖掘(SOLAM)。是建立在多维视图基础之上,基于网络的验证型空间数据挖掘和分析的工具,强调执行效率和对用户命令的及时响应。

2. 解释下列概念

(1) 网格文件

(2) 专题地图

(3) 空间数据仓库

(4) 数字地球

参考答案:略。

3. 简述空间挖掘的意义。

参考答案:空间挖掘通常被称作空间数据挖掘,或者空间数据库的知识发现,它是从空间数据库中抽取隐含的知识、空间关系或非显式地存储在空间数据库中的其他模式,用于理解空间数据、发现数据间的关系。由于大量的空间数据从各种应用中收集而来,收集到的数据远远超过了人脑分析的能力。空间挖掘就是为了满足高效空间数据处理的需要而出现的。

4. 举例说明空间挖掘的意义。

参考答案:略。

5. 简述空间数据的特征。

参考答案:由于空间属性的存在,空间数据具有复杂性的特征,主要表现在:

(1) 空间属性之间的非线性关系;

(2) 空间数据的多尺度特征;

(3) 空间信息的模糊性;

（4）空间维数的增高；

（5）空间数据的缺值现象。

6. 简述空间查询的类型。

参考答案：略。

7. 常用的空间数据索引结构有哪些？

参考答案：常用的空间数据索引结构有：网格文件、四叉树、*R* 树、*k*-D 树。

8. 与传统数据库索引技术相比，空间索引方法具有什么样的特殊性？常用的空间数据索引结构有哪些？

参考答案：略。

9. 基于两个空间实体的位置，空间实体之间的拓扑关系可以概括为哪些种类？

参考答案：两个空间实体之间存在的拓扑关系有：分离、重叠/相交、等价、包含于、覆盖/包含。

10. 假设 A 和 B 是二维空间中的两个空间实体，基于两个空间实体的位置，空间实体之间的拓扑关系可以概括为哪些种？

参考答案：略。

11. 简述空间数据的泛化方法。

参考答案：空间数据的泛化包括空间数据支配泛化和非空间数据支配泛化。空间数据支配泛化做的是基于空间位置的聚类，非空间数据支配泛化根据非空间属性值的相似性做聚类，归纳出高层次的模式或特征。当空间数据（或非空间数据）归纳之后，非空间数据（或空间数据）进行适当的调整，以反映新的空间区域所联系的非空间数据。

12. 简述空间数据支配泛化算法的主要思想。

参考答案：略。

13. 请给出空间规则的概念与表示方法。

参考答案：空间规则是在一定的知识背景下，对数据进行概括和综合，在空间数据库或空间数据仓库中搜索和挖掘规则和规律，得到的以概念树形式给出的高层次的模式或特征。

在空间数据挖掘中有以下三种类型的规则。

（1）空间特性规则：描述数据。

（2）空间判别规则：描述不同种类数据间的差异。

（3）空间关联规则：是两个数据集合之间的关联。

14. 请说出空间关联规则与传统关联规则的关系与区别。

参考答案：略。

15. 简述空间决策树的基本思路。

参考答案：要建造一个空间决策树，首先找到空间或非空间的相关谓词，然后用最相关的谓词来建造树。

16. 请说出空间决策树与一般决策树的关系。

参考答案:略。

17. 常用的空间聚类方法有哪些?

参考答案:常用的空间聚类方法有:基于随机搜索的聚类方法 CLARANS 扩展,大型空间数据库基于距离分布的聚类算法 DBCLASD,BANG 方法,小波聚类,近似值方法。

18. 请列举常用的空间聚类方法,并对这些方法进行比较。

参考答案:略。

19. 简述 SOLAP 的主要任务。

参考答案:空间联机分析处理 SOLAP 是针对特定问题的联机空间数据访问和分析。在空间数据挖掘的早期阶段,SOLAP 工具可以帮助用户分析数据,找到比较重要的变量,发现异常数据和互相影响的变量,帮助用户更好地理解数据,加快知识发现的过程。

20. 请结合 GeoMiner,谈谈一个空间数据挖掘系统应该具备的主要功能与体系结构。

参考答案:略。

PART 2

第二部分

各章授课重点与课时分配

第1章　绪　　论

本章总学时估计在6～9.5学时,教师可根据讲授的对象和总学时计划进行安排。

1.1　数据挖掘技术的产生与发展(1学时)

主要介绍清楚数据挖掘技术的商业需求和技术产生的背景。

1.2　数据挖掘研究的发展趋势(0.5或1学时)

从技术发展角度阐述清数据挖掘技术在研究和应用上将来可能的重点工作。

1.3　数据挖掘概念(1或1.5学时)

从不同角度解释清楚数据挖掘的技术含义。

1.4　数据挖掘技术的分类问题(0.5学时)

介绍清楚不同的分类方法及其指导意义。

1.5　数据挖掘常用的知识表示模式与方法(1.5或2学时)

从宏观上介绍清楚数据挖掘的知识表示模式与方法,不必追求细节,以后章节再展开讲。

1.6　不同数据存储形式下的数据挖掘问题(1或1.5学时)

从宏观上介绍清楚不同数据存储形式下的数据挖掘可能面对的问题与对策,不必追求细节,以后章节再展开讲。

1.7　粗糙集方法及其在数据挖掘中的应用(可选或1学时)

主要介绍清楚粗糙集的相关概念以及和数据挖掘技术的关系。

1.8　数据挖掘的应用分析(0.5或1学时)

主要通过实例介绍清楚数据挖掘技术的应用价值,以激发学生的学习积极性。

第2章　知识发现过程与应用结构

本章总学时估计在5～9学时，教师可根据讲授的对象和总学时计划进行安排。

2.1　知识发现的基本过程(1～1.5学时)

主要是系统化地介绍清楚知识发现的基本过程和主要阶段。对各阶段的功能要给出明确的解释。

2.2　数据库中的知识发现处理过程模型(1.5～2学时)

重点讲述阶梯处理过程模型、螺旋处理模型。对以用户为中心的处理模型、联机KDD模型从用户交互角度阐述它们的必要性和基本思想。支持多数据源多知识模式的KDD处理模型可以根据情况选讲。

2.3　知识发现软件或工具的发展(0.5～1.5学时)

重点讲述清楚知识发现软件或工具发展的3个主要阶段和含义。对教材中给出的KDD系统可以根据情况选讲。

2.4　知识发现项目的过程化管理(0～1学时)

可以根据情况选讲。介绍时应该强调知识发现的过程管理的重要性和必要性。

2.5　数据挖掘语言介绍(2～3学时)

重点介绍清楚数据挖掘语言的种类和思想。对数据挖掘查询语言机器DMQL要讲解清楚它的具体技术，已给学生一个较完整地概念。对其他两类语言可以情况有重点地介绍。

第3章　关联规则挖掘理论和算法

本章总学时估计在9～28学时，对于研究生教学来讲，建议安排15学时以上。教师可根据讲授的对象和总学时计划进行安排。

3.1　基本概念与解决方法(1学时)

重点介绍清楚事务数据库、项目集、支持度、频繁项目集、可信度、关联规则等概念。对关联规则挖掘的两个主要阶段的功能要阐述清楚。

3.2　经典的关联规则挖掘算法分析(3学时)

重点讲解项目集空间理论、Apriori算法。对Apriori算法和对应的关联规则生成算法，要通过实例让学生掌握它解决问题的具体步骤。

3.3　Apriori算法的性能瓶颈问题(1学时)

重点讲述Apriori算法的两个主要性能瓶颈。

3.4　Apriori的改进算法(2～3学时)

对3个改进算法的提出背景、解决的问题等要加以介绍。重点讲解它们的算法的基本思想。假如学时充裕，可以增加实例来说明上述问题。

3.5　对项目集空间理论的发展(2～4学时)

重点介绍清楚这种发展的必要性，由此讲解清楚Close和PF-Tree算法的基本思想。假如学时充裕或者学生基础好，可以通过增加Close和PF-Tree算法的运行实例来说明上述问题。

3.6　项目序列集格空间和它的操作(0～2学时)

这部分内容较难，教师可以根据情况选讲。

3.7　基于项目序列集操作的关联规则挖掘算法(0～2学时)

这部分属于选择内容，教师可以根据情况选讲(假如讲解的话，必须提前讲解3.6节)。

3.8　改善关联规则挖掘质量问题(0～1学时)

这部分属于选择内容，教师可以根据情况选讲。

3.9　约束数据挖掘问题(0～2学时)

这部分属于选择内容，教师可以根据情况选讲。

3.10　时态约束关联规则挖掘(0～2学时)

这部分属于选择内容，教师可以根据情况选讲(假如讲解的话，必须提前讲解3.6节和3.7节)。

3.11　关联规则挖掘中的一些更深入的问题(0～3学时)

这部分属于选择内容，教师可以根据情况选讲(假如讲解的话，建议提前讲解3.9节)。讲解中要注意通俗易懂，有实例说明。

3.12　数量关联规则挖掘方法(0～4学时)

这部分属于选择内容，教师可以根据情况选讲。

第4章　分类方法

本章总学时估计在7～20学时,对于研究生教学来讲,建议安排15学时以上。教师可根据讲授的对象和总学时计划进行安排。

4.1　分类的基本概念与步骤(1学时)

重点介绍清楚分类的概念,对数据分类的两个步骤要阐述清楚。

4.2　基于距离的分类算法(1～2学时)

重点讲述基于距离的分类算法的基本思路,教师可以根据情况选讲 *k*NN 算法。假如学时充裕,可以增加实例来说明上述问题。

4.3　决策树分类方法(3～6学时)

重点介绍清楚决策树分类方法的两个基本步骤,由此讲解清楚 ID3 算法的基本思想。教师可以根据情况选讲 C4.5 算法(假如讲解的话,建议提前讲解 ID3 算法)。假如学时充裕或者学生基础好,可以通过增加 ID3 算法和 C4.5 算法的运行实例来说明上述问题。

4.4　贝叶斯分类(2～4学时)

重点讲解清楚贝叶斯定理以及贝叶斯分类的工作过程,假如学时充裕或者学生基础好,可以通过增加贝叶斯分类的运行实例来说明上述问题。EM 算法属于选讲内容。

4.5　规则归纳(0～6学时)

这部分属于选择内容,教师可以根据情况选讲。

4.6　与分类有关的其他问题(0～1学时)

这部分属于选择内容,教师可以根据情况选讲。

第5章　聚类方法

本章总学时估计在6～12学时，对于研究生教学来讲，建议安排6学时以上。教师可根据讲授的对象和总学时计划进行安排。

5.1　概述(1～2学时)

对聚类方法在数据挖掘中的地位和典型的应用要阐述清楚，重点介绍清楚聚类的概念、聚类的基本分类、距离与相似性度量方法。

5.2　划分聚类方法(1～3学时)

重点介绍清楚划分聚类的主要思想，由此讲解清楚 k-平均算法的基本思想。教师可以根据情况选讲 PAM 算法。

5.3　层次聚类方法(2～3学时)

重点介绍清楚层次聚类的主要思想，由此讲解清楚 AGENS 算法和 DIANA 算法。教师可以根据情况选讲其他层次聚类方法。

5.4　密度聚类方法(2学时)

重点介绍清楚密度聚类的主要思想，由此讲解清楚 DBSCAN 算法。

5.5　其他聚类方法(0～2学时)

这部分属于选择内容，教师可以根据情况选讲。

第 6 章　时间序列和序列模式挖掘

本章总学时估计在 5～16 学时,对于研究生教学来讲,建议安排 5 学时以上。教师可根据讲授的对象和总学时计划进行安排。

6.1　时间序列及其应用(1 学时)

主要介绍时间序列的概念和它的应用。

6.2　时间序列预测的常用方法(1 学时)

重点介绍确定性时间序列预测方法和随机时间序列预测方法。

6.3　基于 ARMA 模型的序列匹配方法(0～2 学时)

这部分属于选择内容,教师可以根据情况选讲。

6.4　基于离散傅里叶变换的时间序列相似性查找(0～2 学时)

这部分属于选择内容,教师可以根据情况选讲。

6.5　基于规范变换的查找方法(0～2 学时)

这部分属于选择内容,教师可以根据情况选讲。

6.6　序列挖掘(1～2 学时)

重点介绍序列挖掘的基本改变,数据源形式和序列模式挖掘的一般步骤。

6.7　AprioriAll 算法(1～2 学时)

重点讲解清楚 AprioriAll 算法的基本思想,假如学时充裕或者学生基础好,可以通过运行实例来说明上述算法。

6.8　AprioriSome 算法(1～2 学时)

重点讲解清楚 AprioriSome 算法的基本思想,假如学时充裕或者学生基础好,可以通过运行实例来说明上述算法。

6.9　GSP 算法(0～2 学时)

可以根据情况选讲 GSP 算法。

第7章　Web挖掘技术

本章总学时估计在6～13学时，对于研究生教学来讲，建议安排6学时以上。教师可根据讲授的对象和总学时计划进行安排。

7.1　Web挖掘的意义(1学时)

主要介绍Web挖掘的意义和它的应用。

7.2　Web挖掘的分类(1学时)

重点介绍Web内容挖掘、Web访问信息挖掘、Web结构挖掘。

7.3　Web挖掘的含义(1学时)

重点介绍Web挖掘的含义，重点区别Web挖掘与信息检索、Web挖掘与信息抽取。

7.4　Web挖掘的数据来源(1学时)

结合Web挖掘应用场景，介绍Web挖掘数据来源。

7.5　Web内容挖掘方法(1～2学时)

重点介绍Web内容挖掘的概念、主要技术和预处理。

7.6　Web访问信息挖掘方法(1～5学时)

重点介绍Web访问信息挖掘的特点和意义，Web访问信息挖掘的数据源以及预处理，Web访问信息挖掘的常用技术和要素组成。假如学时充裕或者学生基础好，可以对Web访问信息挖掘的应用进行介绍。

7.7　Web结构挖掘方法(0～2学时)

可以根据情况选讲Web结构挖掘方法。

第8章　空间挖掘

本章总学时估计在7～12学时,对于研究生教学来讲,建议安排7学时以上。教师可根据讲授的对象和总学时计划进行安排。

8.1　引言(0.5学时)

主要介绍空间数据和空间数据挖掘的一般性概念和应用。

8.2　空间数据概要(1.5～2学时)

讲解清楚空间数据的特征和查询问题;空间数据的主要数据结构等。

8.3　空间数据挖掘基础(0.5学时)

简要介绍空间挖掘需要的基础知识。

8.4　空间统计学(0～0.5学时)

假如需要,简要介绍。

8.5　泛化与特化(1～2学时)

讲解清楚泛化与特化的概念,以及对应的算法。

8.6　空间规则(1学时)

讲解清楚空间规则描述空间实体的结构及它们之间关系的方法、类型以及对应的算法。

8.7　空间分类算法(0.5～1学时)

讲解清楚空间分类方法以及对应的算法。

8.8　空间聚类算法(1～1.5学时)

讲解清楚空间聚类方法,选择性介绍对应的算法。

8.9　空间挖掘的其他问题(0～0.5学时)

简单地介绍空间挖掘对应在线分析、图像等多媒体信息挖掘、可视化等问题及常见的解决思路。

8.10　空间数据挖掘原型系统介绍(0～0.5学时)

假如需要,介绍GeoMiner。

8.11　空间数据挖掘的研究现状(0.5学时)

归纳性介绍。

8.12　空间数据挖掘的研究与发展方向(0.5学时)

归纳性介绍。

8.13　空间数据挖掘与相关学科的关系(0～0.5学时)

假如需要,简单介绍。

8.14　数字地球(0～0.5学时)

假如需要,简单介绍。

PART 3

第三部分

按总学时规划的教学大纲

48 学时的教学大纲（本科生）

【课程名称】

数据挖掘技术（Technology of Data Mining）。

【总学时】

48 学时。

【授课对象】

计算机科学与技术专业本科生。

【先修课程】

数据库原理等。

【课程目的与地位】

数据挖掘技术经过十几年的发展，已经取得一批重要成果，特别是在基本概念、基本原理、基本算法等方面发展的越来越清晰。因此，现在开设此课具备基本的技术条件。本课程以介绍基本概念和基本算法为主，以前沿问题的讨论与探索为辅，其目的是为学生将来研究和学习提供知识储备。

数据处理技术是计算机相关专业培养而设置的重点课程群之一。这个课程群的基础性课程是数据库原理，解决数据处理中的数据表示、关系数据库的管理以及查询等基本问题。数据挖掘作为高级数据处理和分析技术，是这个课程群的高级课程，其目的是通过本课程学习让学生了解信息处理技术的发展方向以及数据挖掘技术本身的概念、原理和方法。

【教材】

毛国君等. 数据挖掘原理与算法. 北京：清华大学出版社，2007.

【主要参考书目】

Jiawei Han，Micheline Kambr. Data Mining：Concepts and Techniques. 影印版. 北京：高等教育出版社，2001.

【教学内容、基本要求及学时分配】

1．第 1 章，绪论（6 学时）。

本章作为绪论，其目的是让学生对数据挖掘技术有一个总体的认识。因此，主要内容是对数据挖掘技术的概念、产生背景技术、发展趋势以及应用等进行提炼和概括。

学时的分配为：

- 数据挖掘技术的产生与发展趋势（2 学时）。
- 数据挖掘技术的分类与知识表示模式（2 学时）。
- 不同数据存储形式下的数据挖掘问题与应用等介绍（2 学时）。

2．第 2 章，知识发现过程与应用结构（5 学时）。

本章对 KDD 过程及其应用模型结构进行阐述，其目的是从系统应用角度给读者一个关于 KDD 设计和实现的技术概括。

学时的分配为：

- 知识发现的基本过程（1 学时）。
- 主要的知识发现处理过程模型介绍（2 学时）。
- 知识发现软件与挖掘语言介绍等（2 学时）。

3. 第 3 章,关联规则挖掘理论和算法(10 学时)。

本章对关联规则挖掘中的概念、方法、算法进行全面的分析和论述。由于关联规则挖掘是数据挖掘技术中研究最早、成果最多、相对比较成熟的分支,因此本章重点在于一些经典理论和算法、热点问题的介绍。

学时的分配为:

- 基本概念与解决方法(1 学时)。
- Apriori 算法(2 学时)。
- Apriori 的改进算法(2 学时)。
- Close 和 FP-tree 算法(2 学时)。
- 数量关联规则挖掘方法介绍(2 学时)。
- 其他的一些高级技术介绍(1 学时)。

4. 第 4 章,分类方法(10 学时)

分类在数据挖掘中是一项非常重要的任务,本章对分类的基本概念与步骤、经典的分类方法以及与分类有关的问题进行了阐述。

学时的分配为:

- 分类的基本概念与步骤(1 学时)。
- 基于距离的分类算法(1 学时)。
- 决策树分类方法(2 学时)。
- 贝叶斯分类(2 学时)。
- 规则归纳有选择性的介绍(2 学时)。
- 与分类有关的其他问题介绍(2 学时)。

5. 第 5 章,聚类方法(9 学时)。

聚类是数据挖掘技术中一个重要内容,内容很多,因此本章主要从基本方法、按划分聚类方法、层次聚类方法和密度聚类方法等进行重点讲解。

学时的分配为:

- 聚类分析的概念、基本方法归纳(2 学时)。
- 基于划分的聚类方法与算法(2 学时)。
- 基于层次的聚类方法与算法(2 学时)。
- 基于密度的聚类方法与算法(2 学时)。
- 其他聚类方法介绍(1 学时)。

6. 第 6 章,时间序列和序列模式挖掘(4 学时,可选)

第 6 章和第 8 章的内容相对较新,可以根据需要,作如下选择:

(1)只选择其中一章,使该章占用 4 学时。

(2)对相应的基本内容做简单介绍,每章各占 2 学时。

本章学时的分配为:

- 时间序列预测的常用方法介绍(1 学时)。
- 序列挖掘的基本方法介绍(1 学时)。
- 时间序列预测的典型算法介绍(1 学时)。
- 序列挖掘的典型算法介绍(1 学时)。

7. 第 7 章 Web 挖掘技术(4 学时)

由于 Web 挖掘是数据挖掘领域崭新的研究分支,所以许多方法具有探索性。因此本章重点是来阐述 Web 挖掘所要解决的主要问题和意义,并选择了一些研究比较集中和相对比较成熟或被认可的技术进行论述。

学时的分配为:

- Web 挖掘的意义、含义、应用、主要方法归纳(2 学时)。
- Web 访问信息挖掘方法与算法介绍(1 学时)。
- 其他 Web 挖掘方法与算法介绍(1 学时)。

8. 第 8 章,空间挖掘(4 学时,可选)

同上所述,第 6 章和第 8 章的内容相对较新,可以根据需要,作如下选择:

(1) 只选择其中一章,使该章占用 4 学时。

(2) 对相应的基本内容做简单介绍,每章各占 2 学时。

本章学时的分配为:

- 空间数据挖掘特点、含义与基础方法介绍(2 学时)
- 空间数据库挖掘的典型算法介绍(2 学时)

32学时的教学大纲(本科生)

【课程名称】

数据挖掘技术(Technology of Data Mining)。

【总学时】

32学时。

【授课对象】

计算机科学与技术专业本科生。

【先修课程】

数据库原理等。

【课程目的与地位】

数据挖掘技术经过十几年的发展,已经取得一批重要成果,特别是在基本概念、基本原理、基本算法等方面发展的越来越清晰。因此,现在开设此课具备基本的技术条件。本课程主要介绍基本概念、基本方法,选择一些典型的数据挖掘算法进行讲解,其目的是为学生将来从事相关工作提供知识储备。

其他信息参考上面的48学时大纲。

【教材】

毛国君等.数据挖掘原理与算法.北京:清华大学出版社,2007.

【主要参考书目】

Jiawei Han,Micheline Kambr. Data Mining:Concepts and Techniques.影印版.北京:高等教育出版社,2001.

【教学内容、基本要求及学时分配】

1. 第1章,绪论(6学时)

本章作为绪论,其目的是让学生对数据挖掘技术有一个总体的认识。因此,主要内容是对数据挖掘技术的概念、产生背景技术、发展趋势以及应用等进行提炼和概括。

学时的分配为:

- 数据挖掘技术的产生与发展趋势(2学时)。
- 数据挖掘技术的分类与知识表示模式(2学时)。
- 不同数据存储形式下的数据挖掘问题与应用等介绍(2学时)。

2. 第2章,知识发现过程与应用结构(5学时)

本章对KDD过程及其应用模型结构进行阐述,其目的是从系统应用角度给读者一个关于KDD设计和实现的技术概括。

学时的分配为:

- 知识发现的基本过程(1学时)。
- 主要的知识发现处理过程模型介绍(2学时)。
- 知识发现软件与挖掘语言介绍等(2学时)。

3. 第3章,关联规则挖掘理论和算法(9学时)

本章对关联规则挖掘中的概念、方法、算法进行全面的分析和论述。由于关联规则挖掘

是数据挖掘技术中研究最早、成果最多、相对比较成熟的分支，因此本章重点在于一些经典理论和算法、热点问题的介绍。

学时的分配为：

- 基本概念与解决方法(2 学时)。
- Apriori 算法(2 学时)。
- Apriori 的改进算法(2 学时)。
- Close 和 FP-tree 算法(2 学时)。
- 其他的一些高级技术介绍(1 学时)。

4．第 4 章，分类方法(6 学时)

分类在数据挖掘中是一项非常重要的任务，本章对分类的基本概念与步骤、经典的分类方法以及与分类有关的问题进行了阐述。

学时的分配为：

- 分类的基本概念与步骤(1 学时)。
- 基于距离的分类算法(1 学时)。
- 决策树分类方法(2 学时)。
- 其他分类方法介绍(2 学时)。

5．第 5 章，聚类方法(6 学时)

聚类是数据挖掘技术中一个重要内容，内容很多，因此本章主要从基本方法、按划分聚类方法、层次聚类方法和密度聚类方法等进行重点讲解。

学时的分配为：

- 聚类分析的概念、基本方法归纳(2 学时)。
- 基于划分的聚类方法与算法(2 学时)。
- 其他聚类方法介绍(2 学时)。

48学时的教学大纲(研究生)

【课程名称】

数据挖掘技术(Technology of Data Mining)。

【总学时】

48学时。

【授课对象】

计算机科学与技术专业研究生。

【先修课程】

数据库原理等。

【课程目的与地位】

数据挖掘技术经过十几年的发展,已经取得一批重要成果,特别是在基本概念、基本原理、基本算法等方面发展的越来越清晰。近年来,数据挖掘及其相关技术已经成为研究生进行科学研究的主要方向之一,而且数据挖掘作为方法可以被许多其他研究领域来使用。因此,在研究生中开设该课程不论是对于学生在校研究和将来工作都具有重要的理论和应用价值。数据处理技术的智能化将成为将来计算机应用的核心技术之一,而传统的数据技术只能解决数据查询等基本问题,因此利用在校的宝贵时间,让研究生对数据挖掘理论与技术有一个全面而正确地认识,是一件非常有意义的工作。

【教材】

毛国君等.数据挖掘原理与算法.北京:清华大学出版社,2007.

【主要参考书目】

Jiawei Han,Micheline Kambr.Data Mining:Concepts and Techniques.影印版.北京:高等教育出版社,2001.

【教学内容、基本要求及学时分配】

1. 第1章,绪论(5学时)

本章作为绪论,其目的是让学生对数据挖掘技术有一个总体的认识。作为研究生教学,应该力求从理论框架和技术发展轨迹的视点,提出问题,给出正确而全面地关于数据挖掘技术的概念、产生背景技术、发展趋势以及应用等方面的概括。

学时的分配为:

- 数据挖掘技术的产生与发展、概念(2学时)。
- 数据挖掘技术的分类与知识表示模式(1学时)。
- 不同数据存储形式下的数据挖掘问题(1学时)。
- 粗糙集方法及其在数据挖掘中的应用(1学时)。

2. 第2章,知识发现过程与应用结构(4学时)

本章立足于从KDD系统的技术构架及其应用模型角度阐述问题,其目的是为研究生将来进行系统研发提供入门性的技术指导。

学时的分配为:

- 知识发现的基本过程(1学时)。

- 主要的知识发现处理过程模型介绍(2 学时)。
- 知识发现软件与挖掘语言介绍(1 学时)。

3. 第 3 章,关联规则挖掘理论和算法(8 学时)

本章对关联规则挖掘中的概念、方法、算法进行全面的分析和论述。由于关联规则挖掘是数据挖掘技术中研究最早、成果最多、相对比较成熟的分支,因此本章重点在于一些经典理论和算法、热点问题的介绍,同时对于研究生来讲要提出一些开放性的问题进行讨论。

学时的分配为:

- 基本概念与解决方法(1 学时)。
- Apriori 算法(1.5 学时)。
- Apriori 的改进算法(1 学时)。
- Close 和 FP-tree 算法(2 学时)。
- 数量关联规则挖掘方法介绍(0.5 学时)。
- 开放性问题讨论,如多维数据的挖掘、多层次概念的发现、约束数据挖掘等(2 学时)。

4. 第 4 章,分类方法(8 学时)

分类在数据挖掘中是一项非常重要的任务,而且应用非常广泛。应该除了对分类的基本概念与步骤、经典的分类方法进行介绍外,要对它的一些最新发展有个讨论。

学时的分配为:

- 分类的基本概念与步骤(1 学时)。
- 基于距离的分类算法(1 学时)。
- 决策树分类方法(1 学时)。
- 贝叶斯分类(1 学时)。
- 规则归纳有选择性的介绍(1 学时)。
- 与分类有关的其他问题介绍(1 学时)。
- 开放性问题讨论,如分类方法的性能评估、集成分类器学习、分类与处理问题等(2 学时)。

5. 第 5 章,聚类方法(8 学时)

聚类分析也是数据挖掘技术中一个重要内容,有很强的应用适应性。因此除了对基本概念与方法进行较细致的讲解外,要重点对划分聚类方法、层次聚类方法和密度聚类方法等进行剖析。也应该设置一些开放性的问题进行讨论。

学时的分配为:

- 聚类分析的概念、基本方法归纳(2 学时)。
- 基于划分的聚类方法与算法(1 学时)。
- 基于层次的聚类方法与算法(1 学时)。
- 基于密度的聚类方法与算法(1 学时)。
- 其他聚类方法介绍(1 学时)。
- 开放性问题讨论,如开放性问题,如聚类与分类方法的区别、模糊聚类问题等(2 学时)。

6. 第 6 章,时间序列和序列模式挖掘(4 学时)

本章主要讲解基本概念和典型算法。

学时的分配为:

- 时间序列预测的常用方法介绍(1 学时)。
- 序列挖掘的基本方法介绍(1 学时)。
- 时间序列预测的典型算法介绍(1 学时)。
- 序列挖掘的典型算法介绍(1 学时)。

7. 第 7 章,Web 挖掘技术(7 学时)

由于 Web 挖掘是数据挖掘领域崭新的研究分支,所以许多方法具有探索性。因此本章重点是来阐述 Web 挖掘所要解决的主要问题和意义,并选择了一些研究比较集中和相对比较成熟或被认可的技术进行论述。教师也应该根据情况选择一些开放性的问题进行讨论。

学时的分配为:

- Web 挖掘的意义、含义、应用、主要方法归纳(2 学时)。
- 典型的 Web 访问信息挖掘方法与算法(1 学时)。
- 典型的 Web 结构挖掘方法与算法(1 学时)。
- Web 内容挖掘中的问题与方法探讨(1 学时)。
- 开放性问题讨论,如开放性问题,如无结构或者半结构数据的挖掘问题、电子商务数据以及个性化网页推荐等应用性问题等(2 学时)。

8. 第 8 章,空间挖掘(4 学时)

本章主要讲解基本概念和典型算法。

学时的分配为:

- 空间数据挖掘特点、含义与基础方法介绍(2 学时)。
- 空间数据库挖掘的典型算法介绍(2 学时)。

P A R T 4

第四部分 样本试卷

样本试卷 1(本科生)

一、(20 分)简明扼要地解释下列概念、并且给出它们对应的英文表达。

1. 数据挖掘。
2. 机器学习。
3. 数据库。
4. 人工智能。
5. 数据仓库。

二、(20 分)KDD 是一个多步骤的处理过程,它一般包含哪些基本阶段?简述各阶段的主要功能。

三、(20 分)对表 t-1 给出的一个事务数据库,跟踪 Apriori 算法生成频繁项目集的过程,其中最小支持度为 50%。

表 t-1　事物数据库

TID	Itemset	TID	Itemset
1	A,B,C,D	3	A,B,C,E
2	B,C,E	4	A,B,C,D

四、(20 分)图 t-1 是使用 ID3 算法在一个数据集上生成的决策树,它用来帮助银行来决定是否发放住房贷款。根据该图回答下列问题:

1. 数据格式至少包含哪些属性?定义一个数据表来满足这种格式要求。
2. 写出该树对应的分类规则。

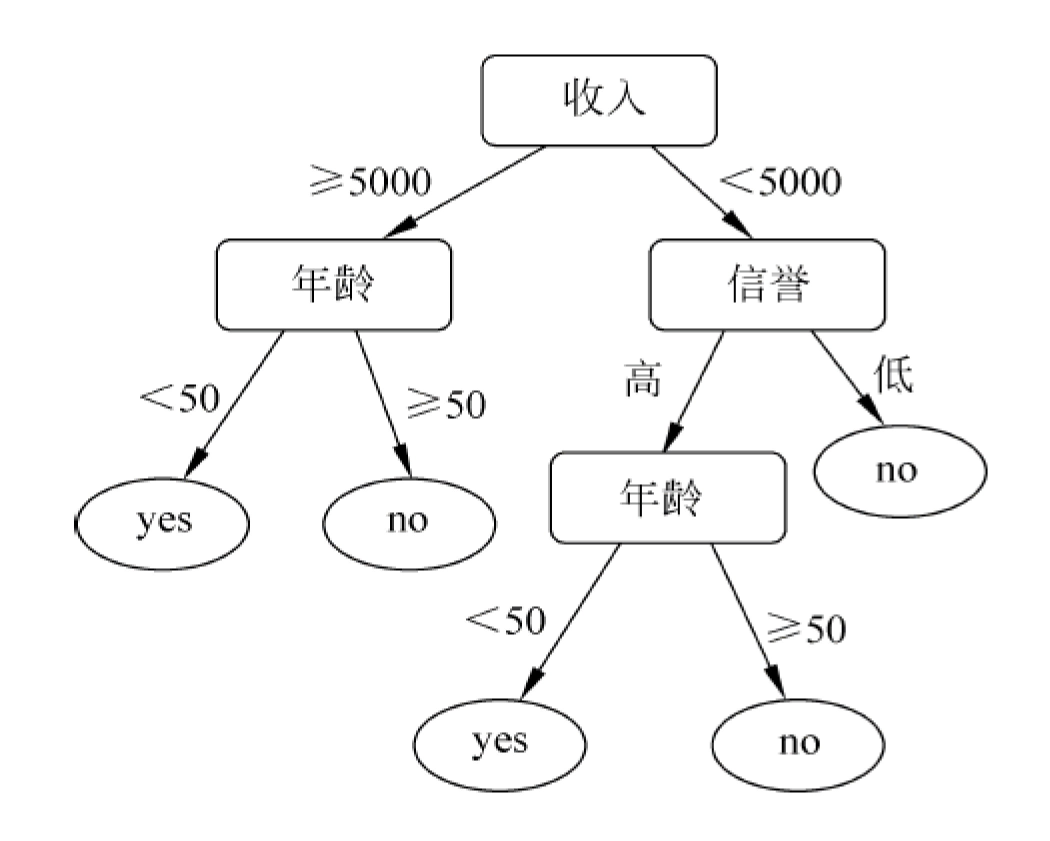


图 t-1　第四题对应的决策树

五、(20 分)对给定的数据 data={a=(1,1), b=(2,1), c=(1,2), d=(2,2), e=(4,3), f=(5,3), g=(4,4),h=(5,4)},实施 k-means(假如 k=2,开始的中心点是 a 和 c)。在计算中心点和距离时参照下面的计算方法:

(1) 中心点采用平均值,如(2,3)和(4,5)的中心点应该是(3,4)。

(2) 距离采用欧氏距离,如(2,3)和(4,5)的距离是 $\sqrt{(2-4)^2+(3-5)^2}=2$。

样本试卷 2(研究生)

一、(20 分)指出下列英文单词或缩写的中文含义,并简单地解释它们。

1. OLAP。
2. Information Retrieval。
3. Bayesian Classification。
4. Data Warehouse。
5. Spatial mining。
6. Vertical Solution of data mining。
7. KDD。
8. Association rule。
9. Clustreing。
10. Data Visualization

二、(15 分)有人说数据库、统计学和人工智能是支撑数据挖掘研究的三个主要基础学科,请说明这种说法的合理性和局限性。

三、(20 分)Apriori 算法是最早的数据挖掘典型算法,针对它回答下列问题:

1. 对表 t-2 给出的一个事务数据库,跟踪 Apriori 算法的生成频繁项目集的过程,其中最小支持度为 50%。

表 t-2 事物数据库

TID	Itemset	TID	Itemset
1	A,B,C,D	4	A,B,D,E
2	B,C,D,E	5	A,B,C,D
3	A,B,C,F		

2. 说明 Apriori 方法存在的主要问题,尝试给出解决这些问题的主要途径或者思想。

四、(10 分)简述 ID3 算法的基本思想、主要问题和改进策略。

五、(15 分)k-means 算法是一种广泛使用的聚类算法。请回答下列问题:

1. 对给定的数据 data={a=(1,1),b=(2,1),c=(3,1),d=(1,3),e=(1,4),f=(2,3),g=(3,2),h=(4,1)},实施 k-means(假如 k=2,假设开始的中心点是 a 和 d)。在计算中心点和距离时参照下面的计算方法:

(1) 中心点采用平均值,如(2,3)和(4,5)的中心点应该是(3,4)。

(2) 距离采用欧氏距离,如(2,3)和(4,5)的距离是 $\sqrt{(2-4)^2+(3-5)^2}=2$。

2. 为了提高算法的效率和适应性,可以从哪些方面进行改进?

六、(20 分)从下列选择一个数据挖掘的研究分支,讨论该分支所要解决的主要问题、面对的挑战性课题,给出解决这些问题或者挑战将要采取的技术性或策略性的设想。

可选择的题目有:

分支一:Web 挖掘。

分支二:空间挖掘。

分支三:多维数据库挖掘。

分支四:时间序列挖掘。

样本试卷1(本科生)的参考答案

一、参考答案

对应的英文术语如下：

1. Data Mining。
2. Machine Learning。
3. Database。
4. Artificial Intelligence。
5. Data Warehous。

二、参考答案

KDD是一个多步骤的处理过程，一般分为问题定义、数据抽取、数据预处理、数据挖掘以及模式评估等基本阶段。

问题定义阶段的功能：确定挖掘目标等。

数据抽取阶段的功能：选取相应的源数据库，并根据要求从数据库中提取相关的数据。

数据预处理阶段的功能：对前一阶段抽取的数据进行再加工，检查数据的完整性及数据的一致性等。

数据挖掘阶段的功能：运用选定的数据挖掘算法，从数据中提取出用户所需要的知识。

知识评估阶段的功能：将KDD系统发现的知识以用户能了解的方式呈现，并且根据需要进行知识评价。如果发现知识和用户挖掘目标不一致，则重复以上阶段以最终获得可用的知识。

三、参考答案

L_1 生成：生成候选集并通过扫描数据库得到它们的支持数，$C_1=\{(A,3),(B,4),(C,4),(D,2),(E,2)\}$；挑选 minsup_count $\geqslant 2$ 的项目集组成1-频繁项目集 $L_1=\{A,B,C,D,E\}$。

L_2 生成：由 L_1 生成2-候选集并通过扫描数据库得到它们的支持数 $C_2=\{(AB,3),(AC,3),(AD,2),(AE,1),(BC,4),(BD,2),(BE,2),(CD,2),(CE,2),(DE,0)\}$；挑选 minsup_count $\geqslant 2$ 的项目集组成2-频繁项目集 $L_2=\{AB,AC,AD,BC,BD,BE,CD,CE\}$。

L_3 生成：由 L_2 生成3-候选集并通过扫描数据库得到它们的支持数 $C_3=\{(ABC,3),(ABD,2),(ACD,2),(BCD,2),(BCE,2)\}$；挑选 minsup_count$\geqslant 2$ 的项目集组成3-频繁项目集 $L_3=\{ABC,ABD,ACD,BCD,BCE\}$。

L_4 生成：由 L_3 生成4-候选集并通过扫描数据库得到它们的支持数 $C_4=\{(ABCD,2)\}$；挑选 minsup_count$\geqslant 2$ 的项目集组成4-频繁项目集 $L_4=\{ABCD\}$。

于是所有的频繁项目集为$\{A,B,C,D,E,AB,AC,AD,BC,BD,BE,CD,CE,ABC,ABD,ACD,BCD,BCE,ABCD\}$。另外很容易得到最大频繁项目集为$\{ABCD,BCE\}$。

四、参考答案

1. T＝<income, age, rate, loan>，其中income代表收入，实型数据；age代表年龄，整型数据；rate代表信誉，是值域为{high, low}的布尔型数据；loan是决策属性，取值为yes表明可以贷款、取值为no表明不能贷款。

2. 使用上面的符号，分类规则可以写成：

If (income≥5000) and (age<50) Then loan=yes;

If (income<5000) and (rate='high') and (age<50) Then loan=yes;

If (income≥5000) and (age≥50) Then loan=no;

If (income<5000) and (rate='low') Then loan=no;

If (income<5000) and (rate='high') and (age≥50) Then loan=no。

五、参考答案

第一次迭代：选取中心点是 a 和 c，分别找到离 a 和 c 最近的对象，并产生两个簇$\{a, b, d\}$、$\{c, e, f, g, h\}$。

对于产生的簇分别计算平均值，得到平均值点。

对于$\{\{a, b, d\}\}$，平均值点为(5/3, 4/3)；

对于$\{c, e, f, g, h\}$，平均值点为(19/5, 16/5)。

第二次迭代：通过平均值调整对象的所在的簇，重新聚类，即将所有点按离平均值(5/3, 4/3)、(19/5, 16/5)最近的原则重新分配。得到 2 个新的簇：$\{a, b, c, d\}$和$\{e, f, g, h\}$。重新计算簇平均值点，得到新的平均值点为(3/2,3/2)、(9/2, 7/2)。

第三次迭代：将所有点按离平均值(3/2,3/2)、(9/2, 7/2)最近的原则重新分配。仍然为$\{a, b, c, d\}$和$\{e, f, g, h\}$，发现准则函数收敛，程序结束。

样本试卷 2(研究生)的参考答案

一、参考答案

1. 在线分析处理。

2. 信息检索。

3. 贝叶斯分类。

4. 数据仓库。

5. 空间挖掘。

6. 数据挖掘的纵向解决方案。

7. 数据库中的知识发现。

8. 关联规则。

9. 聚类。

10. 数据可视化。

上面术语的解释参考教材。

二、参考答案

合理的回答要点:

(1) 任何技术的产生总是有它的技术背景的,数据库、统计学和人工智能的发展导致数据挖掘的技术需求;

(2) 数据库的普及性应用,改变人们利用存储海量数据的方式;

(3) 统计学是任何数据分析的基础,数据挖掘系统的核心模块技术和算法离不开它的支持;

(4) 人工智能等的理论与技术性成果为数据挖掘技术的提出和发展起到了极大地推动作用。

局限的回答要点:

(1) 数据挖掘不能等同于这些技术的叠加;

(2) 数据挖掘有不同的处理问题的思想。

三、参考答案

1. 频繁项目集生成过程如下:

L_1 生成:生成候选集并通过扫描数据库得到它们的支持数,$C_1=\{(A,4),(B,5),(C,4),(D,4),(E,2)\}$;挑选 minsup_count ≥ 3 的项目集组成 1-频繁项目集 $L_1=\{A,B,C,D\}$。

L_2 生成:由 L_1 生成 2-候选集并通过扫描数据库得到它们的支持数 $C_2=\{(AB,4),(AC,3),(AD,2),(BC,4),(BD,4),(CD,3)\}$;挑选 minsup_count ≥ 3 的项目集组成 2-频繁项目集 $L_2=\{AB,AC,AD,BC,BD,CD\}$。

L_3 生成:由 L_2 生成 3-候选集并通过扫描数据库得到它们的支持数 $C_3=\{(ABC,3),(ABD,3),(ACD,2),(BCD,3)\}$;挑选 minsup_count$\geq 3$ 的项目集组成 3-频繁项目集 $L_3=\{ABC,ABD,BCD\}$。

L_4 生成:由 L_3 生成 4-候选集,为空。

于是所有的频繁项目集为$\{A,B,C,D,AB,AC,AD,BC,BD,CD,ABC,ABD,BCD\}$。另外很容易得到最大频繁项目集为$\{ABC,ABD,BCD\}$。

2. Apriori 方法存在的问题及对策,参考教材中相关内容。

四、参考答案

关于 ID3 算法的基本思想、主要问题和改进策略,参考教材中相关内容。

五、参考答案

1. k-means 算法执行过程如下:

第一次迭代:选取中心点是 a 和 d,分别找到离 a 和 d 最近的对象,并产生两个簇 $\{a, b, c, g, h\}$、$\{d, e, f\}$。

对于产生的簇分别计算平均值,得到平均值。

对于 $\{a, b, c, g, h\}$,平均值为(13/5, 6/5);

对于 $\{d, e, f\}$,平均值为(4/3, 10/3)。

第二次迭代:通过平均值调整对象的所在的簇,重新聚类,即将所有点按离平均值(13/5, 6/5)和(4/3,10/3)最近的原则重新分配。得到 2 个新的簇:$\{b, c, g, h\}$ 和 $\{a, d, e, f\}$。重新计算簇平均值点,得到新的平均值点为(9/2, 7/2)和(3, 5/4)。

第三次迭代:将所有点按离平均值(9/2, 7/2)或者(3, 5/4)最近的原则重新分配。仍然为 $\{b, c, g, h\}$ 和 $\{a, d, e, f\}$,发现准则函数收敛,程序结束。

2. 改进 k-means 算法,参考教材中相关内容。

六、参考答案

这是一个综合考查题,教师应该根据学生选择的分支,参考教材中对应部分和相关的研究文献来评测。

读者意见反馈

亲爱的读者：

感谢您一直以来对清华版计算机教材的支持和爱护。为了今后为您提供更优秀的教材，请您抽出宝贵的时间来填写下面的意见反馈表，以便我们更好地对本教材做进一步改进。同时如果您在使用本教材的过程中遇到了什么问题，或者有什么好的建议，也请您来信告诉我们。

地址：北京市海淀区双清路学研大厦 A 座 602 室　计算机与信息分社营销室　收
邮编：100084　电子邮件：jsjjc@tup.tsinghua.edu.cn
电话：010-62770175-4608/4409　邮购电话：010-62786544

教材名称：数据挖掘原理与算法（第二版）教师用书
ISBN：978-7-302-19350-0

个人资料

姓名：________ 年龄：________ 所在院校/专业：________________
文化程度：________ 通信地址：________________________
联系电话：________ 电子信箱：________________________

您使用本书是作为： □指定教材 □选用教材 □辅导教材 □自学教材

您对本书封面设计的满意度：
□很满意 □满意 □一般 □不满意　改进建议________________

您对本书印刷质量的满意度：
□很满意 □满意 □一般 □不满意　改进建议________________

您对本书的总体满意度：
从语言质量角度看　□很满意 □满意 □一般 □不满意
从科技含量角度看　□很满意 □满意 □一般 □不满意

本书最令您满意的是：
□指导明确 □内容充实 □讲解详尽 □实例丰富

您认为本书在哪些地方应进行修改？（可附页）

____

____

您希望本书在哪些方面进行改进？（可附页）

____

____

电子教案支持

敬爱的教师：

为了配合本课程的教学需要，本教材配有配套的电子教案（素材），有需求的教师可以与我们联系，我们将向使用本教材进行教学的教师免费赠送电子教案（素材），希望有助于教学活动的开展。相关信息请拨打电话 010-62776969 或发送电子邮件至 jsjjc@tup.tsinghua.edu.cn 咨询，也可以到清华大学出版社主页（http://www.tup.com.cn 或 http://www.tup.tsinghua.edu.cn）上查询。